edited by
hilary greaves
jacob barrett
david thorstad

ESSAYS ON LONGTERMISM

present action for the distant future

OXFORD

Essays on Longtermism

Essays on Longtermism

Present Action for the Distant Future

Edited by

HILARY GREAVES, JACOB BARRETT,
AND DAVID THORSTAD

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© the several contributors 2025

The moral rights of the authors have been asserted

This is an open access publication, available online and distributed under the
terms of a Creative Commons Attribution-Non Commercial-No Derivatives 4.0
International licence (CC BY-NC-ND 4.0), a copy of which is available at
https://creativecommons.org/licenses/by-nc-nd/4.0/.
Subject to this license, all rights are reserved.

Enquiries concerning reproduction outside the scope of this licence should be sent to
the Rights Department, Oxford University Press, at the address above.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number is on file at the Library of Congress

ISBN 9780192883858

DOI: 10.1093/9780191979972.001.0001

Printed and bound by
CPI Group (UK) Ltd., Croydon, CR0 4YY

Contents

List of Contributors ix

1. Introduction 1
Hilary Greaves, Jacob Barrett, and David Thorstad

PART 1. EVALUATING THE CASE FOR LONGTERMISM

2. The Case for Strong Longtermism 17
Hilary Greaves and William MacAskill
3. Longtermism and Neutrality about More Lives 50
Katie Steele
4. Prudential Longtermism 65
Johan E. Gustafsson and Petra Kosonen
5. Would a World Without Us Be Worse? Clues from Population Axiology 86
Andreas L. Mogensen
6. Longtermism in an Infinite World 105
Christian Tarsney and Hayden Wilkinson
7. Longtermism and the Complaints of Future People 126
Emma J. Curran
8. Against a Moral Duty to Make the Future Go Best 139
Charlotte Franziska Unruh
9. Authenticity, Meaning, and Alienation: Reasons to Care Less about Far-Future People 150
Stefan Riedener

PART 2. PREDICTING AND EVALUATING THE FUTURE

10. What Are the Prospects of Forecasting the Far Future? 171
David Rhys Bernard and Eva Vivalt
11. Taking the Long View: Paleobiological Perspectives on Longtermism 180
Rachell Powell
12. Coping with Myopia 197
Philip Kitcher
13. Shaping Humanity's Longterm Trajectory 211
Toby Ord
14. Longtermism and Cultural Evolution 238
Aron Vallinder

PART 3. ETHICAL PRIORITIES

15. The Hinge of History and the Choice between Patient and Urgent Longtermism 257
Olle Häggström

16. How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role 272
Carl Shulman and Elliott Thornley

17. Longtermist Myopia 295
Amanda Askell and Sven Neth

18. Minimal and Expansive Longtermism 315
Hilary Greaves and Christian Tarsney

19. What Would a Longtermist Society Look Like? 334
Owen Cotton-Barratt and Rose Hadshar

20. Is Extinction Risk Mitigation Uniquely Cost-Effective? Not in Standard Population Models 343
Gustav Alexandrie and Maya Eden

21. Depopulation and Longtermism 362
Michael Geruso and Dean Spears

22. Existential Risk from Power-Seeking AI 383
Joe Carlsmith

23. Deceit and Power: Machine Learning and Misalignment 410
Richard Ngo and Adam Bales

24. The Ethics, Economics, and Demographics of Delaying Aging 428
Kevin Kuruc and David Manley

25. Longtermism and Animals 449
Heather Browning and Walter Veit

PART 4. INSTITUTIONS AND SOCIETY

26. Longtermist Political Philosophy: An Agenda for Future Research 465
Andreas T. Schmidt and Jacob Barrett

27. Retrospective Accountability: A Mechanism for Representing Future Generations 492
Tyler M. John

28. Longtermism and Social Risk-Taking 511
H. Orri Stefánsson

29. The Short-Termism of 'Hard' Economics 529
Ilan Noy and Shakked Noy

30. The Intuitive Appeal of Legal Protection for Future Generations 547
Eric Martínez and Christoph Winter

31. Temporal Distance Reduces Ingroup Favoritism 567
Stefan Schubert, Lucius Caviola, Julian Savulescu, and Nadira S. Faber

Index 581

Contributors

Gustav Alexandrie is a Predoctoral Research Fellow in Economics at the Global Priorities Institute, University of Oxford. His current work focuses on welfare economics, social choice, and decision theory.

Amanda Askell is a Research Scientist at Anthropic, focusing on AI alignment. She has a PhD in Philosophy from NYU and a BPhil in Philosophy from the University of Oxford.

Adam Bales is a Senior Research Fellow at the Global Priorities Institute at the University of Oxford, as well as a Research Fellow at Wolfson College.

Jacob Barrett is an Assistant Professor of Philosophy and Political Science at Vanderbilt University. Before coming to Vanderbilt, he was a Senior Research Fellow in Philosophy at the Global Priorities Institute at the University of Oxford. He works on social, moral, and political philosophy, and especially on questions relating to long-run social reform.

David Rhys Bernard is a PhD candidate at the Paris School of Economics, specializing in impact evaluation. His research focuses on assessing the internal validity of different methods for impact evaluation in both the short and long run, with the aim of improving our ability to estimate long-run effects. He is a Global Priorities Fellow at the Forethought Foundation and was a Fulbright Scholar at UC Berkeley.

Heather Browning is a Lecturer in Philosophy at the University of Southampton. She specializes in animal welfare, ethics, and consciousness. Prior to her current position, she worked as a researcher in animal sentience and welfare at the London School of Economics, as part of the Foundations of Animal Sentience project. She previously completed her PhD in Philosophy at the Australian National University, writing on conceptual and methodological questions in the measurement of animal welfare. Alongside her academic career, Browning has also worked as a zookeeper and animal welfare officer.

Joe Carlsmith is a Senior Research Analyst at Open Philanthropy, where he focuses on existential risk from advanced artificial intelligence. He also writes independently about philosophy and futurism. He has a Doctorate in Philosophy from Oxford University.

Lucius Caviola is a moral psychologist at Harvard University with a PhD from the University of Oxford. His research focuses on judgment and decision-making in the altruistic context. It investigates the psychological tendencies that exacerbate existential risk and prevent us from having a large positive impact on the world. Lucius has published research articles in leading academic journals, such as *Psychological Science, Journal of Personality and Social Psychology, Trends in Cognitive Sciences*, and *Science Advances*, as well as a book about the psychology of effective altruism at Oxford University Press.

Owen Cotton-Barratt is an independent researcher with a DPhil in Mathematics. His research interests centre on finding practical guidance for choosing actions in the face of large uncertainty about the future, and the unpredictability of technological progress. He has published in *Global Policy*, *Risk Analysis*, the *Journal of Philosophy*, and *Health Security*, among others.

Emma Curran is a PhD candidate in Philosophy at the University of Cambridge, working in ethics and metaphysics. In 2023, Emma will be joining the Center for Population-Level Bioethics at Rutgers University—New Brunswick.

Maya Eden is an Associate Professor of Economics at Brandeis University and a Senior Affiliate of the Global Priorities Institute at Oxford University. Her primary research areas are normative economics and macroeconomics.

Nadira Faber is a Research Associate at the Oxford Uehiro Centre for Practical Ethics and a Full Professor at the University of Bremen (Germany), where she leads the Social and Economic Psychology Unit. Nadira Faber conducts empirical research with implications for ethics on topics around moral psychology, helping and altruism, human–animal relationships, cooperation, and group dynamics.

Michael Geruso is an Associate Professor in the Economics Department of the University of Texas at Austin and a Faculty Research Associate at the National Bureau of Economic Research. Geruso also serves as the Assistant Director in UT-Austin's Population Wellbeing Initiative. His work spans topics in health, economic demography, and social welfare evaluation of public policy. Prior to joining UT-Austin, he was a Robert Wood Johnson Postdoctoral Fellow at Harvard University, and earned his PhD at Princeton University.

Hilary Greaves is Professor of Philosophy at the University of Oxford. Her research interests range broadly across ethics, but with a particular focus around issues of axiology and those lying at the interface with economics. Greaves's theoretical work has spanned, among other things, utilitarian aggregation, population axiology, interpersonal comparisons of well-being, moral uncertainty, discounting, and cluelessness. She also has worked on various issues of practical ethics, including healthcare prioritization, population size, global poverty, climate change, artificial intelligence, and existential risk. From 2017 to 2022, Greaves served as Founding Director of the Global Priorities Institute at the University of Oxford.

Johan E. Gustafsson is a Research Scientist at the University of Texas at Austin, a Senior Research Fellow in Philosophy at University of York, and a Docent in Practical Philosophy at the University of Gothenburg and at the Institute for Futures Studies. He is the author of *Money-Pump Arguments* (Cambridge University Press, 2022) and works primarily on the parts of philosophy that relate to the question of what we ought to do, either morally or rationally. His work covers theoretical problems in decision theory, ethics, value theory, free will, personal identity, and political philosophy.

Rose Hadshar is an independent researcher focused on questions related to AI governance. Trained in history at the Universities of Oxford and York, her current interests include forecasting AI development, the regulation of frontier AI systems, and extreme risks from such systems.

Olle Häggström is Professor of Mathematical Statistics at Chalmers University of Technology and member of the Royal Swedish Academy of Sciences. The bulk of his research qualifications are in probability theory, but in recent years he has increasingly reoriented towards work on AI futurology, existential risk, and related topics. He is the author of five books, including *Here Be Dragons: Science, Technology and the Future of Humanity* (Oxford University Press, 2016).

Tyler John is a Programme Officer at Longview Philanthropy, where he leads Longview's work on international policy, priorities research, and academic field-building. Previously, he was a Presidential Fellow at Rutgers University—New Brunswick, where he earned his PhD in Philosophy. Tyler is a Research Affiliate at the Legal Priorities Project and has held fellowships with the Forethought Foundation for Global Priorities Research and the National Institutes of Health. His work can be

found in the *Australasian Journal of Philosophy*, *Economics and Philosophy*, *Ethics*, and the *Oxford Handbook of Consequentialism*.

Philip Kitcher is the John Dewey Professor Emeritus of Philosophy at Columbia University. He is the author of many books, ranging from the philosophy of science to ethics, political philosophy, pragmatism, and philosophy and literature. Among them is *The Seasons Alter*, co-authored with Evelyn Fox Keller, which studies the challenge posed by climate change.

Petra Kosonen is a Postdoctoral Fellow at the Population Wellbeing Initiative at the University of Texas at Austin. Prior to this role, she earned a DPhil in Philosophy from the University of Oxford. Her research focuses on decision theory and normative ethics, and the central theme of her thesis was how we should treat tiny probabilities of vast value.

Kevin Kuruc is a Senior Research Fellow at the Population Wellbeing Initiative of the University of Texas at Austin and a Senior Research Affiliate at the University of Oxford's Global Priorities Institute. Dr. Kuruc is a macroeconomist with expertise in climate, welfare, and population economics. Prior to his current role, he was an Assistant Professor at the University of Oklahoma and received his PhD from the University of Texas at Austin.

William MacAskill is Associate Professor of Philosophy at Oxford University and Senior Research Fellow at the Global Priorities Institute. His research focuses on the fundamentals of effective altruism—the use of evidence and reason to help others as much as possible with our time and money—and on how to act given moral uncertainty. He has published in journals such as *Mind*, *Nous*, and *The Journal of Philosophy*, is the author of the books *Doing Good Better* and *What We Owe The Future*, and is the co-author of *Moral Uncertainty*.

David Manley is Associate Professor of Philosophy at the University of Michigan, Ann Arbor.

Eric Martinez is a PhD candidate in Cognitive Science at MIT and a Research Fellow at the Legal Priorities Project. Eric holds a JD from Harvard Law and is admitted to the Massachusetts bar.

Andreas L. Mogensen is a Senior Research Fellow at Oxford's Global Priorities Institute, and was previously a Tutorial Fellow at Oxford's Jesus College and an Examination Fellow at All Souls College. His research interests are primarily in normative and applied ethics, as well as decision theory, focusing primarily on issues related to ethics and the long-term future. He has publications in journals including the *Journal of Philosophy*, *Philosophy and Phenomenological Research*, and *Philosophers' Imprint*.

Sven Neth is an Assistant Professor of Philosophy at the University of Pittsburgh. His research is about decision theory and formal epistemology with special focus on non-ideal agents. His recent publications include 'A Dilemma for Solomonoff Prediction' (in *Philosophy of Science*) and 'Accuracy and Infinity: A Dilemma for Subjective Bayesians' (*Synthese*, co-authored with Mikayla Kelley). He is currently the recipient of a Josephine De Karman Fellowship and was supported by a Global Priorities Fellowship from the Forethought Foundation.

Richard Ngo is a Research Scientist at OpenAI working on AI governance and AI alignment. His research focuses on the intersection of technical and strategic considerations related to the development of AGI, including understanding potential alignment problems, developing effective AGI governance regimes, and forecasting the effects of AGI on the world. He previously worked as a Research Engineer at DeepMind, and has a background in both computer science and philosophy (as a former PhD candidate in the Philosophy of Machine Learning at the University of Cambridge).

Ilan Noy is the Chair in the Economics of Disasters and Climate Change—Te Āwhionukurangi, at Te Herenga Waka—Victoria University of Wellington. His research and teaching focus on the economic

aspects of natural hazards, disasters, and climate change, and other related topics in environmental, development, and international economics. He is also the founding Editor-in-Chief of the journal *Economics of Disasters and Climate Change*, published by Springer, and has consulted for the World Bank, the Asian and Inter-American Development Banks, OECD, UNDRR, the IMF, and ASEAN.

Shakked Noy is a PhD student in Economics at MIT. He is interested in labour economics, behavioural economics, and political economy; his previous research has examined labour market institutions and the effects of artificial intelligence on labour markets.

Toby Ord is a Senior Research Fellow in Philosophy at Oxford University. His work focuses on the big picture questions facing humanity. What are the most important issues of our time? How can we best address them? Toby's earlier work explored the ethics of global health and global poverty, eventually leading him to co-found the effective altruism movement. His current research is on the longterm future of humanity and the risks which threaten to destroy our entire potential. His book, *The Precipice*, concludes that safeguarding our future is among the most pressing and neglected issues we face.

Rachell Powell is Director of the Center for Philosophy and History of Science and Professor in the Department of Philosophy at Boston University. Her recent books include *Contingency and Convergence: Toward a Cosmic Biology of Body and Mind* (MIT Press) and *The Evolution of Moral Progress: A Biocultural Theory* (Oxford University Press, with Allen Buchanan). She has held fellowships at the Oxford Uehiro Centre for Practical Ethics, the National Humanities Center, the American Council of Learned Societies, the Konrad Lorenz Institute, the National Evolutionary Synthesis Center, the Berman Institute for Bioethics, and the Berlin School of Mind and Brain.

Stefan Riedener is Associate Professor in Practical Philosophy at the University of Bergen. He earned his DPhil from the University of Oxford and previously worked as a Senior Teaching and Research Assistant at the University of Zurich. His main research interests include moral emotions, partiality, and special relationships, as well as normative uncertainty and the ethics of effective altruism.

Julian Savulescu is the Chen Su Lan Professor in Medical Ethics at the National University of Singapore, where he directs the Centre for Biomedial Ethics. An award-winning ethicist and moral philosopher, he trained in neuroscience, medicine, and philosophy, going on to hold the Uehiro Chair in Practical Ethics (2002) at the University of Oxford, where he founded the Oxford Uehiro Centre for Practical Ethics in 2003, before moving to NUS in 2022.

Andreas T Schmidt is Associate Professor of Political Philosophy at the Faculty of Philosophy and the Centre for PPE at the University of Groningen. His research interests include socio-political freedom, consequentialism, distributive justice, egalitarianism, behavioural public policy, public health ethics, and longtermism. He has published in journals such as *American Political Science Review*, *Ethics*, *Philosophy and Public Affairs*, and *American Journal of Bioethics*. In his current project on longtermist political philosophy, he explores what normative institutional principles we should endorse when extending our time horizon from the short- to the medium- and long-term future.

Stefan Schubert is a researcher at the Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, working in the intersection of moral psychology and philosophy. His research focuses on psychological questions of relevance to effective altruism, such as why our altruistic actions are often ineffective, and why we don't invest more in safeguarding our common future.

Carl Shulman is an Advisor to the Open Philanthropy Project and a Research Associate of the Future of Humanity Institute at Oxford University. His research focuses on modelling emerging technologies in artificial intelligence and biotechnology, and prioritizing philanthropic efforts to mitigate associated global catastrophic risks.

Dean Spears is an Associate Professor in the Economics Department and Population Research Center of the University of Texas, where he is Director of the Population Wellbeing Initiative. Dean's research is funded by a population scientist career grant from the National Institutes of Health. He is also Executive Director of r.i.c.e., a research and advocacy non-profit in India, and co-author of *Where India Goes*.

Katie Steele is a Professor in the School of Philosophy at the Australian National University. Her research interests span rational choice and inference, and ethics and public policy. A central research question is what counts as a justified decision, given the stakes, under conditions of uncertainty. Her recent co-authored book *Beyond Uncertainty: Reasoning with Unknown Possibilities* (2021, Cambridge University Press) explores whether and how a decision maker can combat limited 'awareness' about the very possibilities relevant to her choices. In other recent co-authored work she examines whether the proper treatment of moral uncertainty in decision-making is similar to that of empirical uncertainty.

H. Orri Stefánsson is a Professor of Practical Philosophy at Stockholm University, Pro Futura Scientia Fellow at the Swedish Collegium for Advanced Study, and Researcher at the Institute for Futures Studies. His main areas of research are decision theory and ethics. He is currently working on issues related to policy choices in situations of severe uncertainty, with a focus on climate policy, and also on the ethics of fatality risk.

Christian Tarsney is a Research Fellow with the Population Wellbeing Initiative at the University of Texas at Austin. He works mainly on topics in ethics and decision theory, including 'extreme risks' (small probabilities of astronomically good or bad outcomes), decision-making under normative uncertainty, and population ethics. He has also done work in philosophy of time—in particular, why we care more about the future than the past and whether it's rational to do so.

Elliott Thornley is a Postdoctoral Research Fellow at the Global Priorities Institute. His research is in population ethics, decision theory, and AI safety.

David Thorstad is Assistant Professor of Philosophy at Vanderbilt University. Thorstad's work focuses on global priorities research, bounded rationality, and the epistemology of inquiry.

Charlotte Franziska Unruh is an Early Career Research Fellow at the Institute for Ethics in AI, at the Faculty of Philosophy, University of Oxford, and a Research Fellow at Corpus Christi College, Oxford. Charlotte has research interests in the philosophy of harm, the ethics of computing, and the ethics of future generations.

Aron Vallinder is currently an independent researcher pursuing various topics in global priorities research, including the application of cultural evolution to thinking about the long-term future. He previously worked as a Research Fellow at the Forethought Foundation for Global Priorities Research and holds a PhD in Philosophy from the London School of Economics.

Walter Veit is a Postdoctoral Research Associate at the Department of Philosophy, University of Bristol, and an external member of the Munich Center for Mathematical Philosophy at the Ludwig Maximilian University of Munich. He has published on a broad range of topics, but he currently works primarily on (i) the science and ethics of neurodiversity (human, animal, and AI minds), and (ii) the major evolutionary transitions (biological, cognitive, technological, and cultural).

Eva Vivalt is an Assistant Professor in the Department of Economics at the University of Toronto. Dr. Vivalt's main research interests are in investigating stumbling blocks to evidence-based policy decisions, including methodological issues, how evidence is interpreted, and the use of forecasting. Dr. Vivalt is also a principal investigator on several guaranteed income RCTs and has other interests in

labour economics, development, and global priorities research. Dr. Vivalt is a co-founder of the Social Science Prediction Platform, a platform to coordinate the collection of forecasts of research results.

Hayden Wilkinson is a Junior Research Fellow at Wolfson College as well as a Research Fellow at the Global Priorities Institute, both within the University of Oxford. He completed his PhD at the Australian National University in 2021, during which he received a Fulbright scholarship and was a visiting researcher at Princeton University. As a philosopher, he works primarily in normative ethics and decision theory, with a particular focus on topics of global priorities research.

Christoph Winter is Assistant Professor of Law at the Instituto Tecnológico Autónomo de México (ITAM), Research Associate in Psychology at Harvard University, and Director of the Legal Priorities Project.

1
Introduction

Hilary Greaves, Jacob Barrett, and David Thorstad

In the grand expanse of history, our own lives occupy just a tiny speck of time. But the effects of our actions might be a different matter. Those effects could, at least in principle, stretch on down the millennia, affecting the course of that history long after we are gone. How significant is this idea for the decisions we make today?

A cluster of ideas going under the label 'longtermism' hold that considerations of the far future—on timescales of thousands, millions, or even billions of years—are highly significant for today's decision-making. There are several elements to this line of thought. First is the idea that the far future *matters*: that a given episode of suffering, for example, is not very significantly less important than an otherwise identical episode of present suffering, merely on grounds of temporal location. Second is the idea that there are actions we can take today that meaningfully *affect* the course of the far future: its balance, for instance, of happiness over suffering, or the number and/or welfare capacity of far-future welfare subjects. Third, and closely related, is the idea that we today have sufficiently good *epistemic access* to relevant matters concerning our potential effects on the far future. Fourth is the view that the nature of our relationship to these possibilities (assuming they exist at all) is such as to ground strong *reasons*, perhaps even a moral imperative, to act in ways that hold more promise for positively steering the course of the far future.

Besides longtermism, a closely related idea is that of *existential risk*. A number of scholars have raised concerns that certain recent and emerging technologies, if not managed with extreme caution, ground worryingly high risks of either the extinction of the human species within (say) a century or two, or of some other similarly enormous and irreversible catastrophe (Rees 2003; Posner 2004; Häggström 2016; Ord 2020). These same scholars, however, usually also urge that the requisite caution is possible: that there are things that we could do today that would meaningfully reduce these risks, perhaps drastically. Since, in the absence of such a catastrophe, humanity can plausibly expect to survive for (at least) many more millennia, mitigation of such existential risks is one category of plausible intervention to positively affect the course of the far future. However, the idea of longtermism is broader than that of existential risk mitigation, in that, at least in principle, 'longtermism' also encompasses other ways of affecting the far future, besides prevention of existential catastrophe.

In the present volume, a group of eminent scholars from a range of academic disciplines takes on the theme of longtermism, from both sympathetic and critical perspectives. Some of the articles articulate arguments for or against the *truth* of some thesis in the vicinity of the ideas sketched above. Others take as starting point the assumption that considerations of the far future are important, and contribute to working out the implications of that point

Hilary Greaves, Jacob Barrett, and David Thorstad, *Introduction* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Hilary Greaves, Jacob Barrett, and David Thorstad 2025.
DOI: 10.1093/9780191979972.003.0001

of view (for example, working out which interventions seem most strongly recommended, relative to an aim of improving the course of the far future).

Contributors to this volume (where they use the terms at all) define the terms 'longtermism' and 'longtermist' in different ways. All are at least loosely related to the general idea that 'positively influencing the long-term future is a key moral priority of our time' (MacAskill 2022: 4). But under and around this umbrella, potentially important distinctions abound. *Weak* longtermist theses (like the one just quoted) assert only that the far future is *important*; *strong* longtermist theses make comparative claims, to the effect that considerations of the far future are in some specified sense the *most* important. *Axiological* longtermist theses concern the connections between the far future on the one hand, and on the other hand, which actions lead to *better vs. worse outcomes* than which others. *Deontic* longtermist theses concern not (or not directly) the question of what is better or worse, but rather the question of what a given agent, in a given decision situation, *ought or is morally obligated to do*. (While these two issues are of course connected, non-consequentialist accounts of moral obligation hold that they are also importantly distinct.) Some authors use 'longtermism' to denote a *thesis*, while others use the same term to denote an *evaluative perspective* (for example, a perspective that embodies a zero rate of pure time preference, or one according to which considerations of the far future are very important). The adjective 'longtermist', similarly, denotes sometimes a *property of persons* (that of believing a given thesis, adopting a particular perspective or engaging in a particular project), and sometimes a *property of perspectives or projects*. In the present state of the discussion, this flexibility is helpful, so we have made no attempt to regiment the use of the terms 'longtermism' and 'longtermist' throughout the present volume. Nor will we attempt to state any canonical definitions in this introduction. On this terminological point, each essay must be taken on its own terms.

Part 1: Evaluating the Case for Longtermism

The essays in Part 1 focus on articulating and evaluating arguments for and against the truth of some 'longtermist' thesis. As above, it is worth distinguishing between axiological and deontic longtermist theses. The first five chapters examine the case for axiological longtermism, while the subsequent three focus on deontic claims.

In the first main chapter, 'The Case for Strong Longtermism', Hilary Greaves and William MacAskill articulate an argument (building on Beckstead (2013)) for the axiological strong longtermist claim that 'In the most important decision situations facing agents today, (i) Every option that is near-best overall is near-best for the far future [, and] (ii) every option that is near-best overall delivers much larger benefits in the far future than in the near future'.

The argument articulated by Greaves and MacAskill centres around cost-effectiveness estimates for interventions aimed at mitigating existential risks. The risks these authors consider include both risks of human extinction and risks of other 'existential catastrophes' that are primarily matters of low *average welfare* in the far future, rather than matters of bringing forward the date of human extinction.

The argument is simplest if one assumes expectational total utilitarianism: the thesis that one option is (*ex ante*) better than another iff the first corresponds to a higher expectation

value of total welfare than the second. However, Greaves and MacAskill argue, similar conclusions also follow if one departs from expected total utilitarianism in various ways, including on issues of risk aversion, population ethics, and the expected value of the continuation of human civilisation. They also argue that the argument is not undermined either by the thought that we are substantially clueless about the far future and/or our ability to affect it, nor by the thought that we can make only a tiny difference to the probability of the far future taking this course rather than that. Finally, they outline an argument for a deontic version of the strong longtermist thesis, concerning which options a given agent *ought to choose*, rather than (as above) the axiological question which options are *best or near-best*.

The next four chapters all take up, in various ways, the issue of how sensitive arguments around longtermism are to controversial questions in population ethics, or in axiology more generally.

Greaves and MacAskill's chapter argued that longtermism does not depend on a controversial 'totalist' population ethics specifically on the grounds that some longtermist interventions are matters of improving average future welfare without significantly changing the numbers of future people. In 'Longtermism and Neutrality about More Lives', Katie Steele pursues a complementary approach, examining more carefully the question of what a plausible non-totalist theory of population ethics would say about interventions that significantly increase the size of the all-time human population (by delaying extinction). It is often assumed that a theory of population ethics incorporating 'neutrality'—roughly, the idea that adding an extra person with a life worth living does not, in itself, make a state of affairs either better or worse—will tend to hold also that reducing risks of premature extinction is similarly 'neutral'. In that case, one might think, given a choice between relieving more immediate suffering (by, for example, tackling global poverty today) and reducing risks of premature extinction, morally one ought to relieve the immediate suffering. However, Steele argues that this intuitive-seeming reasoning leads one astray. As has long been recognised (Broome 2004, 2005), neutrality is 'greedy', in the sense that a composite change consisting of a neutral change and a change for the better is often itself neutral. Steele argues that because of this 'greediness', theories of population ethics incorporating neutrality in fact imply that one is morally *permitted* (though not required) to pursue reductions in the chance of premature human extinction in preference to more immediate benefits.

In 'Prudential Longtermism', Johan E. Gustafsson and Petra Kosonen consider the case for prudential longtermism, defined to hold of a given person *S* if and only if 'our acts' overall influence on the expected prudential value for *S* is mainly determined by the effects of these acts in the far future'. Gustafsson and Kosonen argue that because of the possibility of radical life extension technology being developed soon enough to (radically) extend the lives of present people, prudential longtermism is true of present people. Because of this, they suggest, longtermism *simpliciter*, specified here as the thesis that 'our acts' expected influence on the expected value of the world is mainly determined by their effects in the far future', is true on a wide variety of ethical theories. This includes both fairly radical person-affecting approaches to population ethics (for example, approaches that accord moral status, relative to a given time, only to people who are alive at the time in question), and common-sense approaches that recommend strongly prioritising the interests of one's friends and family.

Discussions of longtermism often assume with little argument that near-term human extinction (for example, in the present century) would be enormously bad. But many have doubts about that claim, for reasons relating to human suffering and/or the effects of human activity on the welfare of other animals and the biosphere more generally. In 'Would a World Without Us Be Worse? Clues from Population Axiology', Andreas L. Mogensen undertakes a study of this question from the perspective of a variety of different theories of value. Theories considered include negative utilitarianism, theories incorporating a 'procreation asymmetry', total and critical level utilitarianism, average utilitarianism and variable value theories, prioritarianism, egalitarianism, perfectionism, and conservatism. Careful analysis, Mogensen argues, reveals that in many of these cases, the implications for the value of the continuation of human existence are otherwise than is often more casually assumed.

Standard arguments for longtermism assume, for simplicity, that there are only finitely many welfare subjects. In this context, one can specify a simple totalist axiology, according to which the goodness of a given world is that world's total welfare. However, it is not clear how to extend that axiology, or those arguments, to account for the possibility that the world might contain infinitely many welfare subjects. In 'Longtermism in an Infinite World', Tarsney and Wilkinson consider the implications of this challenge for an ethical perspective that seeks to take far-future welfare seriously.

The next three chapters take up specifically the issue of *deontic* longtermism: rather than examining the significance (or otherwise) of the far future in determining *how well our available options should be expected make the world go*, their focus is on the significance (or otherwise) of the far future in determining what we *ought, or are morally obligated, to do*, from non-consequentialist perspectives. All three are sympathetic to views on which the deontic significance of far-future considerations is severely limited.

In 'Longtermism and the Complaints of Future People', Emma J. Curran develops an anti-aggregationist objection to deontic longtermism. On an anti-aggregationist account, smaller benefits and harms (or 'complaints') are not always to be aggregated across persons, and therefore there is less tendency (compared to an aggregationist account) for an option to be justified on the basis of small benefits to many persons. For example, if you can save one person from dying, or relieve many from headaches, anti-aggregationists usually hold that you ought to save the one, no matter how many headaches are at stake. But actions that benefit distant future people characteristically have only a very small probability of benefiting them, whereas at least some actions have a much higher probability of benefitting present and/or nearer-future people. This means that deontic longtermism is *prima facie* less plausible on an anti-aggregationist picture. Curran elaborates this objection as it arises on both *ex ante* and *ex post* versions of anti-aggregationism. She then addresses the concern that one might take this objection to constitute a *reductio* of anti-aggregationism (rather than an argument against deontic longtermism), and considers what implications anti-aggregationism might have for prioritisation between different ways of trying to affect the very long run.

In 'Against a Moral Duty to Make the Future Go Best', Charlotte Franziska Unruh considers a 'deontic strong longtermist' thesis according to which one ought to act so as to near-maximise far-future value, at least when this does not require excessive sacrifice or violate deontological constraints (as defended in Greaves and MacAskill's contribution to this volume). Unruh argues that this thesis presupposes a mistaken picture of the nature

of duties of beneficence. Beneficence, on Unruh's account, is an imperfect duty that only requires us to benefit others to a sufficient extent, enough of the time. On Unruh's favoured analysis, beneficence therefore does not imply any obligation to (near-)maximise benefits or to make the future go (near-)best. This might suggest that making the far future go best is permissible but not required on non-consequentialist views. But Unruh rejects this claim too, suggesting that we have special duties in relation to future people that can outweigh beneficence and render maximising benefits for future people impermissible. For example, we may be morally required to undo threats of harm to future people that we have already put in motion, even in cases where other actions would benefit future people to a greater degree.

In 'Authenticity, Meaning, and Alienation: Reasons to Care Less about Far-Future People', Stefan Riedener takes up the question of whether or not present-day agents morally ought to accord the same weight to welfare gains and losses accruing to present vs. far-future people, for the purposes of guiding decisions. Riedener defends a negative answer. His argument centres around the fact that far-future people will plausibly have lives, and sources of welfare, that are very alien to us today. We are not in a good position to genuinely *appreciate* the welfare increases in question, in the sense that we do not, at any level of detail, understand what they consist in. Consequently, Riedener suggests, while we can (for carefully chosen longtermist interventions) have reasonably good grounds for believing that a given intervention will increase far-future welfare somehow, we cannot properly appreciate those welfare increases, in comparison to more immediate welfare increases that other interventions could bring about. Riedener argues that because of this, (i) acting in a way that is very strongly motivated by prospects of increasing far-future welfare is problematically inauthentic, and (ii) far-future welfare increases do much less to add meaning to a present-day agent's life than present-day or near-future welfare increases do.

Part 2: Predicting and Evaluating the Future

Part 2 of this volume focuses on the challenges involved in making sufficiently good predictions to guide attempts to influence the course of the far future, and in affecting the course it takes in a sufficiently persistent way for intervening on the far future to be significantly beneficial.

The first two chapters in this section directly take up the question of how well we can predict the far future.

In 'What Are the Prospects of Forecasting the Far Future?', David Rhys Bernard and Eva Vivalt explore existing empirical evidence bearing on our ability to reliably forecast the effects of our actions on the distant future. Noting that we have limited direct evidence about the success of very long-term predictions, they approach this issue by examining the extent to which the predictive accuracy of shorter-term forecasts decays over time. Bernard and Vivalt restrict their focus to causal forecasts about the impact of some intervention on some future variable, as opposed to state or conditional forecasts that respectively concern unconditional and conditional values of future variables. They identify 29 existing studies evaluating causal forecasts, which they argue tentatively support the existence of a decay effect. However, they ultimately recommend epistemic humility about whether or not we can

reliably make very long-term forecasts, on the grounds that we simply have little evidence about this either way.

In 'Taking the Long View: Paleobiological Perspectives on Longtermism', Rachell Powell draws on evolutionary theory to outline an argument that we cannot reliably make predictions on timescales of millions or billions of years. Although earlier thinkers believed in a teleological march toward predefined outcomes, and some still believe that evolution tends toward greater complexity, Powell argues that there is little theoretical support for such ideas. More generally, she argues that we cannot identify causal laws governing evolution that support predictions on evolutionary timelines, nor uniformities that support long-term extrapolations from past trends. Instead, the course of macroevolutionary history is highly contingent, with previous developments constraining later ones and apparent instances of convergent evolution nearly always arising due to shared developmental pathways. This suggests that attempts to predictably shape the long-term future are doomed to fail—with one important caveat. The evolution of sentient beings generally, let alone humans with cumulative culture, is highly contingent; if we were to go extinct, beings like us would be very unlikely to re-evolve. The very contingency that prevents us from predicting how the long term will unfold therefore also allows us to predict that our extinction would likely have irreversible effects, and this may lend support to the common longtermist priority of mitigating extinction risks.

The remaining three papers in Part 2 focus on how we might beneficially influence the far future, especially in light of the predictive difficulties of doing so.

In 'Coping with Myopia', Philip Kitcher considers how we ought to deal with long-term problems given an inability to forecast very far into the future. Focusing on the case study of climate change, Kitcher begins by laying out his preferred moral methodology, on which moral progress occurs through well-informed deliberation among all affected parties aiming at a mutually agreeable solution. He then notes that climate change and other long-term problems raise two key challenges for this methodology: that we cannot include future people in our collective deliberations, and that we cannot reliably predict the far future. These challenges are connected. While we might try to resolve the first by employing individuals who understand the plight of future people and are under role obligations to serve as their representatives, the second challenge suggests that such individuals may not exist. Kitcher's solution appeals to the idea that while we cannot reliably predict the *far* future, we may be able to predict the effects of climate change on *near*-future generations reasonably well. He therefore proposes that we appeal to representatives of near-term generations, while also preserving the conditions for near-future generations to carry on their own collective deliberations about how to deal with problems in their near future. In this way, we can act as stewards for far-future generations through a chain that connects each generation to deliberators in subsequent generations, each link of which involves deliberators aiming both to benefit near-future generations and to keep the chain going.

In 'Shaping Humanity's Longterm Trajectory', Toby Ord presents a framework for modelling longterm trajectories of humanity, with the aim of understanding and comparing different ways of changing the longterm trajectory. Ord considers four types of trajectory shifts. *Advancements* advance humanity's progress by making each point in our future history occur temporally earlier by some fixed amount. *Speed-ups* permanently increase the speed of progress at all future times. *Gains* permanently improve the values of all future states by some fixed absolute quantity, while *enhancements* permanently improve the

values of all future states by a fixed fraction of the value they would otherwise have. Using a simple formal model, Ord analyses the conditions under which trajectory shifts of each of these types are likely to be beneficial, and the comparative values of different types of shifts under various assumptions about the default human trajectory.

The field of *cultural evolution* studies the phenomenon by which 'cultural traits' such as ideas, technologies, and values spread and decline, via a process structurally analogous to that of biological evolution. In a process of cultural evolution, a cultural trait that is either more likely to lead to the survival and growth of a group that has that trait, or more likely to be copied and taken up by additional individuals, tends to increase in prevalence over time; cultural traits with the opposite properties go into decline. In 'Longtermism and Cultural Evolution', Aron Vallinder explores what lessons the longtermist project might be able to learn from existing studies of cultural evolution. For example, one might hope for insights about the extent to which the far-future changes one's intervention seeks to bring about seem likely to be supported vs. undermined by processes of cultural evolution, and whether some otherwise promising intervention is unnecessary on the grounds that cultural evolution can be expected to bring about the change in question even without active intervention.

Part 3: Ethical Priorities

The essays in Part 3 take up questions about which types of interventions are most strongly recommended by a goal of improving the course of the far future, as well as the extent to which including concern for the far future changes the decisions that we would make given a concern only with the near term. The first two essays focus on mitigation of existential risks, the usual central focus of contemporary concern for the far future.

According to the *hinge of history hypothesis*, the present century or so is uniquely important for the course of the long-run future. This view is suggested, for example, by the conjunction of the following two thoughts. First, humanity currently faces very significant and unprecedented existential threats from new and emerging technologies. Second, this situation (the 'time of perils') is likely to be temporary, in the sense that if humanity survives these threats for a century or so, the dangers will be very significantly less thereafter. According to this line of thought, it is of critical importance now to do whatever it takes to address these technological threats. This is a version of 'urgent longtermism', which recommends taking certain fairly specific 'object-level' actions now, in the service of safeguarding the far future. Others—so-called 'patient longtermists'—agree that carefully selected actions available today can deliver significant value for the long-run future, but hold that the recommended actions are often matters of general saving and capability building, so as better to empower future decision-makers to navigate whatever challenges the further future presents. In 'The Hinge of History and the Choice between Patient and Urgent Longtermism', Olle Häggström responds to recent criticism of the hinge of history hypothesis due to Will MacAskill (2022), and correspondingly argues in favour of an urgent over a patient approach to longtermism.

Existential risk mitigation is often motivated by appeal to consideration of the course of the very far future, and the potentially enormous number of lives it might hold. But, as we see in many of the other contributions to this volume, *this* route to motivating existential risk

mitigation requires a host of controversial assumptions concerning value, predictability, and the nature of moral obligation. In 'How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role', Carl Shulman and Elliott Thornley argue that these considerations (while perhaps sound) are not necessary in order to make the case that governments should spend more on existential risk mitigation. Via cost-benefit analysis of feasible efforts to reduce near-term existential risks, they argue that existential risk mitigation comes out as cost-effective under standard methods for valuing statistical lives, *even if we only count lives saved in the near term*.

The next three essays continue the investigation of how far-reaching and/or revisionary the practical implications of longtermism are.

In 'Longtermist Myopia', Amanda Askell and Sven Neth argue for a claim that is in some sense a generalisation of Shulman and Thornley's point. Askell and Neth's core claim is that agents who have a zero rate of pure time preference should nonetheless often behave myopically: they should, that is, behave similarly to agents who have a positive rate of pure time preference. Askell and Neth begin by illustrating some senses in which such agents should not behave myopically: they should be more willing than discounting myopic agents to delay action in order to seek information about the future consequences of their actions, and they should place greater emphasis on acts such as existential risk mitigation which preserve option value, leaving potentially valuable acts open for future choice. However, Askell and Neth also discuss four factors pushing even agents with a zero rate of pure time preference to behave myopically: (i) *Causal diffusion*, understood as the tendency for the causal consequences of acts to decay over time; (ii) *Epistemic diffusion*, understood as the tendency for the predictability of causal consequences to decay over time; (iii) M*oral uncertainty* about the correct rate of temporal discounting; and (iv) The *optimism-pessimism dilemma*. According to this last, if the future is likely to be good, then there is less room to improve it, whereas if the future is likely to be bad, then even if in principle there is great room for improvement, in practice improvements may be unlikely to last. Askell and Neth conclude that a zero rate of pure time preference presents in practice a less radical deviation from common moral intuitions than one might initially expect.

Standard arguments for strong longtermism (e.g. Greaves and MacAskill, this volume) rest crucially on cost-effectiveness analyses for projects of existential risk mitigation. Conditional on the correctness of these analyses, it does seem highly likely that the best interventions available to agents today, whatever they are, are all primarily matters of improving the course of the far future. But this is not to say that there exist any, still less that there exist *many*, ways of improving the course of the far future that are not matters of existential risk mitigation. In 'Minimal and Expansive Longtermism', Hilary Greaves and Christian Tarsney call attention to this gap between the ('minimal') thesis that existential risk mitigation is enormously important on the one hand, and the ('expansive') idea that there are *many and varied* promising ways to improve the course of the far future, going significantly beyond matters of existential risk. They also address the related question of whether the far future is an important consideration in a *wide variety* of decision situations vs. only in rather special (albeit especially important) decision situations.

A common objection to temporal impartiality at the societal level is that such impartiality would be extremely demanding on the present generation. A classic version of this worry is that present people might be required to reinvest almost everything they produce to further enrich the future (Arrow 1999). In the context of significant concern about

existential risk, the concern might alternatively be that at least when such risks are present (see, for example, the contributions to this volume by Häggström, Carlsmith, and Ngo and Bales), temporal impartiality would require the present generation to devote almost all its resources to mitigating existential risks. In either case, the worry is that the resulting recommendation makes the prospects for quality of life in the present very bleak, perhaps so much so as to undermine the plausibility of the principles of temporal impartiality that seem to threaten the practical conclusions in question. Greaves and Tarsney's contribution to this volume also makes some remarks on this theme. In 'What Would a Longtermist Society Look Like?', Owen Cotton-Barratt and Rose Hadshar undertake a more focused discussion of this issue, addressing the question of the extent to which such 'demanding' practical conclusions follow from temporal impartiality in the first place. To a very significant extent, Cotton-Barratt and Hadshar suggest, the arguably objectionable conclusions do not follow: the main thing that follows, according to them, is only that the intrinsic value of the welfare of present people is significantly exceeded by its instrumental value.

The next two essays take up questions about the course of the long-run future, and the prospects for influencing it, through the lens of (respectively) economic and demographic models of population. In 'Is Extinction Risk Mitigation Uniquely Cost-Effective? Not in Standard Population Models', Gustav Alexandrie and Maya Eden explore what economic models of population growth imply about the cost-effectiveness of different interventions. If we assume that the value of the future roughly corresponds to the size of the future population, then mitigating risks of outright extinction might seem more cost-effective than anything else. However, Alexandrie and Eden note that this reasoning depends on two empirical assumptions: first, that non-extinction changes to the current population size lack significant effects on long-run population levels, and, second, that the most cost-effective way to increase long-run population levels is to save lives. They then argue that both assumptions are questionable. Alexandrie and Eden thus suggest that those concerned with promoting long-run value would be wise to pay careful attention to the effects of a given catastrophe or intervention on long-run population size, over and above their more immediate effects. For example, catastrophes that 'only' cause deaths may have different long-term effects than those that also destroy capital.

Arguments for longtermism often appeal to the idea that the future may hold vast numbers of humans who are yet to be born. In 'Depopulation and Longtermism', Michael Geruso and Dean Spears review demographic evidence which has led many demographers to conclude that this is not the most likely scenario. On a range of projections, human fertility will soon fall below the rate needed to replace current populations, with many demographers projecting a falling human population at least out to the year 2300. On some models, the size of the total future human population may be rather small: for example, Geruso and Spears review one model on which only 30 billion humans not currently alive today will ever be born. They note that a diminishing human population would lead to diminished economic outputs, with three consequences that might be especially important for the course of the long-run future: decreased specialisation and gains from trade; decreased innovation; and a decreased societal ability to bear fixed costs, including the costs of mitigating existential risks.

The next three essays focus on the relevance of specific emerging technologies to the course of the long-run future, focusing in particular on existential risks from advanced artificial intelligence (AI) and on life-extending technologies.

Consider first the case of AI. Concerns about adverse long-run outcomes from advanced AI can be grouped into three categories. First, one might worry about advanced AI strongly enabling bad actors, for example totalitarian dictators. Second, one might worry about structural disruption to economic and social systems leading to bad outcomes for human welfare, if not successfully managed: for example, massive increases in income inequality and/or unemployment as a result of drastically increased automation of labour. Third, one might worry about futures in which one or more AI systems itself has relevantly agent-like properties, and effectively pursues goals of its own in ways that are incompatible with high welfare. Scholars of artificial intelligence are increasingly becoming concerned about possibilities in this third category (Russell 2019). The next two essays, by Joe Carlsmith ('Existential Risk from Power-Seeking AI') and by Richard Ngo and Adam Bales ('Deceit and Power: Machine Learning and Misalignment'), develop lines of argument in this vein. Both focus on the possibility that agent-like behaviour in sufficiently powerful AI, with relevant abilities significantly exceeding those of humans, might lead to the AI in question seeking and gaining power, and then deploying said power in ways that are incompatible with a flourishing future.

Carlsmith discusses the risks posed by agential systems capable of planning and strategic awareness ('APS systems'). Carlsmith argues that it will likely become both possible and extremely tempting to deploy APS systems by the year 2070. However, Carlsmith argues, it will be difficult to ensure that deployed APS systems are not disposed to seek power over humans on any inputs they will in fact receive. Carlsmith argues that the deployment of APS systems disposed to this form of power-seeking would be quite likely to cause a great deal of damage, up to and including the permanent disempowerment of all (or nearly all) humans.

Like Carlsmith, Ngo and Bales explore concerns that artificial systems might destroy or disempower humanity. In particular, they focus on one component of such arguments: the claim that artificial systems will act in deceptive ways and seek power. Ngo and Bales argue that if advanced systems are developed via deep learning then it's plausible that such behaviours will arise. They see this as a concerning possibility that could potentially play a role in leading to human disempowerment.

Turn next to life extension. In 'The Ethics, Economics, and Demographics of Delaying Aging', Kevin Kuruc and David Manley examine the case, from an impartial point of view, for research into delaying the process of aging in humans. The benefits of life extension, they argue, would be very extensive. Longer lives would, for example, lead to a higher proportion of the population being of working age at a given time. In addition, for many occupations (in which operating in the light of decades of personal experience has high value), longer lives would lead to a higher proportion of each individual's career being spent in the especially productive years that lie between training and cognitive decline. Kuruc and Manley also argue that a wide variety of approaches to population ethics seem likely to converge on the conclusion that delaying aging would be a good thing because of the intrinsic benefits that each individual would enjoy from being able to live longer.

Finally in Part 3, as our discussion to this point reveals, many discussions of longtermism are framed in terms of future *humans*. This is often claimed to be merely for the sake of simplicity of exposition. In 'Longtermism and Animals', however, Heather Browning and Walter Veit argue that existing arguments in fact neglect considerations of non-human animals in ways that lead the arguments seriously astray. Key to their argument is the fact that

non-human animals massively outnumber humans today, and the claim that we should expect them to vastly outnumber humans in the future too. Browning and Veit argue that there is good reason to worry that the animals of the future will experience great amounts of suffering, suggesting that their suffering will be a major source of disvalue in the long run. Browning and Veit thus suggest that those aiming to improve the course of the far future should focus more on interventions targeting animal welfare, which may change either the number or the quality of lives animals lead in the future.

Part 4: Institutions and Society

The essays in Part 4 concern the relationship between society today and the far future. The first three essays take up questions about the design and reform of political institutions and policies.

In 'Longtermist Political Philosophy: An Agenda for Future Research', Andreas T Schmidt and Jacob Barrett lay out longtermist political philosophy as a research field by exploring the case for, and the implications of, 'institutional longtermism'. Institutional longtermism is the view that, when evaluating institutions, we should give significant weight to their very long-term effects. Whereas some discussions of longtermism focus on what individuals should do, Schmidt and Barrett suggest that those attracted to individual longtermism might also favour institutional longtermism, given institutions' large, broad, and long-term effects. Furthermore, institutional longtermism may be able to sidestep several standard objections to individual longtermism, concerning collective action problems, decision-theoretic fanaticism, demandingness, and partiality. However, whereas the standard case for longtermism appeals to the duty of beneficence, political philosophers tend to focus more on values like justice, equality, freedom, legitimacy, and democracy. While such values might initially seem to conflict with institutional longtermism, Schmidt and Barrett argue that these conflicts are less clear-cut upon closer inspection, and that some of these values might even provide independent support for institutional longtermism. Throughout, Schmidt and Barrett's primary aim is to set out central questions in longtermist political philosophy, rather than to definitively answer them; they end with a list of further research questions that are also suggested by the lines of thought in their essay.

In 'Retrospective Accountability: A Mechanism for Representing Future Generations', Tyler M. John tackles the problem of political short-termism: the tendency of political institutions to give undue priority to the near term rather than the long term. John traces this problem to, among other things, a lack of accountability mechanisms incentivising politicians to cater to the interests of future generations. Although we cannot solve this problem by simply giving future people the ability to vote in present-day democracies, John suggests that we may nevertheless be able to incentivise individuals to promote the interests of future people through retrospective accountability mechanisms. More concretely, John proposes a 'futures assembly', in which randomly selected individuals are tasked with promoting the interests of future generations and incentivised to do so via retrospective liability. The core idea is that the pensions of members of this assembly could be partly determined by how well they do at promoting the interests of future generations, as could be evaluated, say, 30 years after their terms in office. John elaborates the institutional details of this proposal, and explores various theoretical issues it raises—for example, whether such mechanisms

would lead the members of this body to promote the interests of individuals in the very distant future, or only those living in the relatively near future.

In 'Longtermism and Social Risk-Taking', H. Orri Stefánsson explores the connection between taking a long-term perspective and incurring social risks. Stefánsson supposes that social planners might reasonably be risk- or loss-averse. Ordinarily, such aversions should render social planners less willing to undertake risky policy experiments. However, Stefánsson argues that if social planners take a long-term perspective that judges social policies in light of their prediction about the long-term future, this should render them significantly more willing to incur risks. His basic idea is that, under certain conditions, evaluating social gambles in combination with other similar gambles or while factoring in other risks should make risk- or loss-averse agents significantly more willing to take risks than when evaluating gambles in isolation. And when we take a long-term perspective, Stefánsson suggests, we must evaluate gambles in combination with the future gambles we expect to take, and future risks we expect to face. Stefánsson then considers various complications concerning, for example, whether we expect the future to be better or worse than the present, and the extent to which the conditions his arguments assume hold in the real world.

In addition to questions about political institutions and policies, longtermism also raises questions about the need to reform academia. In 'The Short-Termism of "Hard" Economics', Ilan Noy and Shakked Noy argue that a bias towards methodological hardness within academic economics has hampered the ability of economists to investigate issues relating to the far future. Noy and Noy discuss three types of norms structuring academic economics: norms of typology which classify research into thematic types, norms of exclusion which classify some research as inadmissible, and norms of omission which mark out some types of admissible research for fewer professional rewards. Noy and Noy argue that these norms have led to inflexible methodological approaches, and thus to the neglect of important research topics. Noy and Noy describe and discuss the limitations of existing academic economic research in three areas that are important from a long-term perspective: conceptual frameworks for long-term decision-making, climate change, and artificial intelligence. Noy and Noy make the case that academic economists could usefully provide, for example, alternative theoretical frameworks or empirical strategies for normative and positive work on far-future considerations. They conclude with some practical recommendations for building a research community focused on these issues within academic economics.

The final two chapters in this volume present and discuss empirical results concerning existing attitudes to the long-run future.

In 'The Intuitive Appeal of Legal Protection for Future Generations', Eric Martínez and Christoph Winter survey a range of recent empirical evidence suggesting that both legal experts and lay people believe that the law can and should be used to protect future people much more than it currently does. Their evidence includes four surveys, respectively of 500 legal academics at non-US universities in the English-speaking world, 600 US law professors, 1,000 lay adults in the US, and 3,000 lay adults across ten countries. Together, this evidence suggests that 'legal longtermism'—the view that the law should be used to protect far future people—is far more intuitive than might otherwise be supposed. However, Martínez and Winter note that the survey evidence also suggests that most people do not believe that future people matter *just as muc*h as present people, and that law professors are not especially optimistic about the prospects for using the law to protect *very* distant future people.

Finally, Martínez and Winter explore whether this evidence provides normative support for legal longtermism and/or strategic guidance for agents seeking to bring it about that the law better protects future people.

Finally, in 'Temporal Distance Reduces Ingroup Favoritism', Stefan Schubert, Lucius Caviola, Julian Savulescu, and Nadira S. Faber present results from psychological studies examining correlations between temporal distance and ingroup/outgroup distinctions, in the context of charitable giving and related tasks. The primary ingroup/outgroup distinction examined is that of nationality: the study participants are US Americans, and they are asked questions about how they would prioritise charitable donations between charities supporting other US Americans on the one hand, and charities supporting more needy people elsewhere in the world on the other hand. The general finding is that US Americans are much more inclined to favour compatriots over foreigners when the beneficiaries are presently existing people than they are when the beneficiaries are in the further future: as time horizons increase, geopolitical ingroup favouritism declines. In addition, the authors find, temporal and geopolitical forms of partiality are correlated: people who are more inclined to favour their compatriots over foreigners are (statistically) more likely to also favour present over future people, and vice versa. The authors argue that the upshot of these findings is that there is a psychological connection between temporal partiality and other forms of partiality.

References

Arrow, K. (1999), 'Inter-generational Equity and the Rate of Discount in Long-Term Social Investment', in M. Sertel (ed.), *Contemporary Economic Issues* (Palgrave Macmillan), 89–102.

Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future*, PhD thesis, Rutgers University.

Broome, J. (2004), *Weighing Lives* (Oxford University Press).

Broome, J. (2005), 'Should We Value Population?', in *Journal of Political Philosophy* 13/4: 399–413.

Häggström, O. (2016), *Here Be Dragons: Technology and the Future of Humanity* (Oxford University Press).

MacAskill, W. (2022), *What We Owe the Future* (Basic Books).

Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Oxford University Press).

Posner, R. (2004), *Catastrophe: Risk and Response* (Oxford University Press).

Rees, M. (2003), *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in this Century—On Earth and Beyond* (Basic Books).

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Books).

PART 1

EVALUATING THE CASE FOR LONGTERMISM

2

The Case for Strong Longtermism

Hilary Greaves and William MacAskill

1 Introduction

A striking fact about the history of civilisation is just how early we are in it. There are 5,000 years of recorded history behind us, but how many years are still to come? If we merely last as long as the typical mammalian species, we still have over 200,000 years to go (Barnosky et al. 2011); there could be a further 1 billion years until the Earth is no longer habitable for humans (Wolf and Toon 2015); and trillions of years until the last conventional star formations (Adams and Laughlin 1999:34). Even on the most conservative of these timelines, we have progressed through a tiny fraction of history. If humanity's saga were a novel, we would be on the very first page.

Normally, we pay scant attention to this fact. Political discussions are normally centred around the here and now, focused on the latest scandal or the next election. When a pundit takes a 'long-term' view, they talk about the next five or ten years. With the exceptions of climate change and nuclear waste, we essentially never think about how our actions today might influence civilisation hundreds or thousands of years hence.

We believe that this neglect of the very long-run future is a serious mistake. An alternative perspective is given by *longtermism*, according to which we should be particularly concerned with ensuring that the far future goes well (MacAskill, 2022). In this essay we go further, arguing for *strong longtermism:* the view that impact on the far future is the *most* important feature of our actions today. We will defend both axiological and deontic versions of this thesis.

Humanity, today, is like an imprudent teenager. The most important feature of the most important decisions that a teenager makes, like what subject to study at university and how diligently to study, is not the enjoyment they will get in the short term, but how those decisions will affect the rest of their life.

The structure of the essay is as follows. Section 2 sets out more precisely the thesis we will primarily defend: axiological strong longtermism (ASL). This thesis states that, in the most important decision situations facing agents today, (i) every option that is near-best overall is near-best for the far future, and (ii) every option that is near-best overall delivers much larger benefits in the far future than in the near future.

We primarily focus on the decision situation of a society deciding how to spend its resources. We use the cost-effectiveness of anti-malarial bednet distribution as an approximate upper bound on attainable near-future benefits per unit of spending. Towards establishing a lower bound on the highest attainable far-future expected benefits, section 3 argues that there is, in expectation, a vast number of sentient beings in the future of human-originating civilisation. Section 4 then argues, by way of examples involving existential risk, that the project of trying to beneficially influence the course of the far future is sufficiently

Hilary Greaves and William MacAskill, *The Case for Strong Longtermism* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Hilary Greaves and William MacAskill 2025.
DOI: 10.1093/9780191979972.003.0002

tractable for ASL(i) and ASL(ii) to be true of the above decision situation. Section 5 argues that the same claims and arguments apply equally to an individual deciding how to spend resources, and an individual choosing a career. We claim these collectively constitute the most important decision situations facing agents today, so that ASL follows.

The remainder of the essay explores objections and extensions to our argument.

Section 6 argues that the case for ASL is robust to several plausible variations in axiology, concerning risk aversion, priority to the worse off, and population ethics. Section 7 addresses the concern that we are clueless about the very long-run effects of our actions. Section 8 addresses the concern that our argument turns problematically on tiny probabilities of enormous payoffs.

Section 9 turns to deontic strong longtermism. We outline an argument to the effect that, according to any plausible non-consequentialist moral theory, our discussion of ASL also suffices to establish an analogous deontic thesis. Section 10 summarises.

The argument in this essay has some precedent in the literature. Nick Bostrom (2003) has argued that total utilitarianism implies we should maximise the chance that humanity ultimately settles space. Nick Beckstead (2013) argues, from a somewhat broader set of assumptions, that 'what matters most' is that we do what's best for humanity's long-term trajectory. In this essay, we make the argument for strong longtermism more rigorous, and we show that it follows from a much broader set of empirical, moral, and decision-theoretic views. In addition, our argument in favour of *deontic* strong longtermism is novel.

We believe that strong longtermism is of the utmost importance: that if society came to adopt the views we defend in this essay, much of what we prioritise in the world today would change.

2 Precisifying strong longtermism

2.1 Axiological strong longtermism

Strong longtermism could be made precise in a variety of ways. First, since we do not assume consequentialism, we must distinguish between axiological and deontic claims. Let axiological (resp., deontic) strong longtermism be the thesis that far-future effects are the most important determinant of the value of our options (resp., of what we ought to do).

It remains imprecise what 'most important determinant' means. Taking the axiological case first, in this essay we consider the following more precise thesis:[1]

> **Axiological strong longtermism (ASL)**: In the most important decision situations facing agents today,
>
> (i) Every option that is near-best overall is near-best for the far future.
>
> (ii) Every option that is near-best overall delivers much larger benefits in the far future than in the near future.

[1] We discuss deontic strong longtermism in section 9.

Where condition (i) holds, one can identify the near-best options by focusing in the first instance only on far-future effects. If (as we believe, but will not argue here) the analogous statement regarding near-future effects is not also true, that supplies one sense in which far-future effects are 'the most important'. Where condition (ii) holds, the evaluation of near-best options is primarily driven by far-future effects. That supplies another such sense.

In sections 3–5, we will argue that clauses (i) and (ii) of ASL hold of particular decision situations: those of a society deciding how to spend money with no restrictions as to 'cause area', an individual making the analogous decision, and individual career choice. Because these decision situations have particularly great significance for the well-being of both present and future sentient beings, we claim, they are the most important situations faced by agents today. Therefore, strong longtermism follows, even if ASL(i) and (ii) do not hold of any other decision situations.

Throughout, 'the far future' means everything from some time t onwards, where t is a surprisingly long time from the point of decision (say, 100 years). 'The near future' means the time from the point of decision until time t. We will interpret both 'near-best overall' and 'near-best for the far future' in terms of proportional distance from zero benefit to the maximum available benefit, and 'much larger' in terms of a multiplicative factor.

As we intend it, ASL is not directly concerned with the *objective* value of options and their *actual* effects. Rather, terms like 'near-best' and 'benefits' relate to the *ex ante* value of those options, given the information available at the time of decision, and their *prospects* for affecting the near or far future. *Ex ante* value may be expected value, but the statement of ASL does not presuppose this.

Since it refers to 'benefits', ASL makes sense only relative to a status quo option: benefits are increases in value relative to the status quo. As above, our primary examples will be cases of deciding how to spend some resource (either money or time). For concreteness, we will then take the status quo to be a situation in which the resources in question are simply wasted. However, other plausible choices of status quo would be unlikely to significantly affect our argument, and the argument does not require that the status quo be special in any deep sense.

ASL makes only comparative claims. We do not claim, and nor do we believe, that options cannot deliver large benefits without being near-best for the far future, or that available near-future benefits are small in any absolute sense. Our claim is rather that available benefits for the far future are many times larger even than this.

2.2 Benefit ratio and ASL

Our argument for ASL proceeds via the intermediate claim that the following property holds of the decision situations in question:

> **Benefit ratio (BR):** The highest far-future *ex ante* benefits that are attainable without net near-future harm are many times greater than the highest attainable near-future *ex ante* benefits.

We prove in the Appendix that if BR holds of a given decision situation, then (first) so does ASL(ii), and (second) ASL(i) holds of a certain restriction of that decision situation. (The restriction involves removing any options that do net expected near-future harm; this restriction is innocuous in the context of our argument.)

Evaluating BR, and hence ASL, requires quantitative analysis; any particular quantitative analysis requires strong evaluative assumptions. To this end, we will *temporarily* make a particular, plausible but controversial, set of evaluative assumptions. Section 6, however, shows that various plausible ways of relaxing these assumptions leave the basic argument intact. One controversial assumption that may be essential, concerning the treatment of very small probabilities, is discussed in section 8.

The inessential assumptions in question include the following. First, we will identify the *ex ante* value of an option with its expected value: the probability-weighted average of the possible values it might result in. Second, we will identify value with total welfare: that is, we will assume a total utilitarian axiology. Third, and a near-corollary of the latter, we will assume time-separability. The latter allows us to separately define near-future and far-future benefits: overall value is then simply the sum of near-future value and far-future value, where these in turn depend only on near-future (resp., far-future) effects.

For a rough upper bound on near-future expected benefits in the context of a society spending money, we consider the distribution of long-lasting insecticide-treated bednets in malarial regions, which saves a life on average for around US$4,000. Each $100 therefore saves on average 0.025 lives in the near future (GiveWell 2020a).[2]

We cannot argue that this is the action with the very largest near-future benefits. In particular, though it seems hard to beat this cost-effectiveness level via any intervention that is backed by rigorous evidence, it might be possible to achieve higher short-term *expected* benefits via some substantially more speculative route.[3] A full examination of the case for strong longtermism would involve investigation of this, and the corresponding sensitivity analysis. However, even quite large upward adjustments to the figure we use here would leave our argument largely unaffected.

We emphasise that we are not considering the long-run knock-on effects of bednet distribution. It is possible, for all we say in this essay, that bednet distribution is the best way of making the far future go well, though we think this unlikely.[4]

We will argue in section 4 that, for a society's decision about how to spend its resources, the lower bound on attainable far-future expected benefits is many times higher than this upper bound for near-future expected benefits, and therefore BR holds of this decision situation. Section 5 discusses related decision situations facing individuals.

[2] Following GiveWell (2018), we will assume that the short-term benefits of the interventions that do the most short-run good would scale proportionately even if very large amounts of money were spent.

[3] A sister organisation to GiveWell, Open Philanthropy, has tried hard to find human-centric interventions that have more short-term impact, and has struggled (Berger 2019). There might be more cost-effective interventions focused on preventing the suffering of animals living in factory farms (Bollard 2016). We leave this aside in order to avoid getting into issues of interspecies comparisons; again, there is a corresponding need for sensitivity analysis.

[4] It would amount to a 'surprising and suspicious convergence' between near-future and far-future optimisation (Lewis 2016).

3 The size of the future

There is, in expectation, a vast number of lives in the future of human civilisation.[5] Any estimate of just how 'vast' is of course approximate. Nonetheless, we will argue, existing work supports estimates that are sufficiently large for our argument to go through.

There are several techniques one can use for estimating the expected number of future beings. Let us start with the question of the expected *duration* of humanity's future existence, temporarily setting aside questions of how large the population might be at any future time.

First, one might use evidence regarding the age of our species to ground judgements on the annual risk of extinction from natural causes, and extrapolate from there. Given that *Homo sapiens* has existed for over 200,000 years, Snyder-Beattie, Ord, and Bonsall (2019:2) thereby estimate that the expected future lifespan of humanity is at least 87,000 years, as far as natural causes of extinction are concerned.

Second, one might undertake *reference class forecasting* (Kahneman and Lovallo 1993; Flyvbjerg 2008). Here, the lifespans of other sufficiently similar species serve as benchmarks. Estimates of the average lifespan of mammalian species (resp., hominins) are between 0.5 and 6 million years (resp., around 1 million years) (Snyder-Beattie et al. 2019:6). Thus reference class forecasting, naively applied, suggests at least 1 million years for the expected future duration of humanity.

Both of these estimates, however, ignore the fact that humans today are highly atypical. Humanity today is significantly better equipped to survive extinction-level threats either than other species are, or than our own species was in the past, thanks to a combination of technological capabilities and geographical diversity. Therefore a range of substantially higher benchmarks is also relevant: for instance, the frequency of mass extinction events (1 in every 30–100 million years (Snyder-Beattie et al. 2019:7)), and the time over which the Earth remains habitable for humans (around 1 billion years (Adams 2008:34)).

The above figures concern the expected *duration* of humanity's future. Since we are interested in the expected *number* of future beings, we also need to consider population size.

We again consider several benchmarks. First, the UN Department of Economic and Social Affairs (2019:6) projects that the global population will plateau at around 11 billion people by the year 2100. Second, the large majority of estimates of the Earth's 'carrying capacity'—that is, its long-run sustainable population, based on relatively conservative assumptions about future technological progress—are over 5 billion, and sometimes substantially higher (Cohen 1995:342; Bergh and Rietveld 2004:197). Third, for predicting the further future, we might extrapolate from the historical trend of human population increasing over time. Given this trend, it is at least plausible that continued technological advances will enable an even larger future population up to some *much* higher plateau point (say, 1 trillion), even if we cannot currently foresee the concrete details of how that might happen (Simon 1998).

Importantly, it is the *expected* number of future beings, not the median, that is relevant for our purposes. In addition to the possibility of numbers like the higher benchmarks indicated above, it is of course also possible that the future duration and/or population size

[5] We will use 'human' to refer both to *Homo sapiens* and to whatever descendants with at least comparable moral status we may have, even if those descendants are a different species, and even if they are non-biological.

of humanity are much *smaller*.[6] However, the effects of these possibilities on the expected number are highly asymmetric. Even a 50% credence that the number of future beings will be *zero* would decrease the expected number by only a factor of two. In contrast, a credence as small as 1% that the future will contain, for example, 1 trillion beings per century for 100 million years (rather than 10 billion per century for 1 million years) increases the expected number by a factor of 100.

We must also consider two more radical possibilities that, while very uncertain, could greatly increase the duration and future population sizes of humanity. The first concerns space settlement. There are currently no known obstacles to the viability of space settlement, and some scientific investigations suggesting its feasibility using known science (Armstrong and Sandberg 2013; Beckstead 2014). If humanity lives not only on Earth but also on other planets—in our own Solar System, elsewhere in the Milky Way, or in other galaxies too—then terrestrial constraints on future population size disappear, and astronomically larger populations become possible. Even if we only settle the Solar System, civilisation would have over 5 billion years until the end of the main sequence lifetime of the Sun (Sackmann, Boothroyd, and Kraemer 1993:462; Schröder and Smith 2008:162), and we would have access to over 2 billion times as much sunlight power as we do now (Stix 2002:6; Sarbu and Sebarchievici 2017:16). If we are able to widely settle the rest of the Milky Way, then we could access well over 250 million rocky habitable-zone planets (Bryson et al. 2021:22), each of which has the potential to support trillions of lives over the course of their sun's lifetime. Moreover, an interstellar civilisation could survive until the end of the stelliferous era, on the order of 10 trillion years hence (Adams and Laughlin 1999). If we consider possible settlement of the billions of other galaxies accessible to us, the numbers get dramatically larger again.

The second radical possibility is that of digital sentience: that is, conscious artificial intelligence (AI). The leading theories of philosophy of mind support the idea that consciousness is not essentially biological, and could be instantiated digitally (Lewis 1980; Chalmers 1996: ch. 9). And the dramatic progress in computing and AI over just the past 70 years should give us reason to think that if so, digital sentience could well be a reality in the future. It is also plausible that such beings would have at least comparable moral status to humans (Liao 2020), so that they count for the purposes of the arguments in this essay.[7]

Consideration of digital sentience should increase our estimates of the expected number of future beings considerably, in two ways. First, it makes interstellar travel much easier: it is easier to sustain digital than biological beings during very long-distance space travel (Sandberg 2014:453). Second, digital sentience could dramatically increase the number of beings who could live around one star: digital agents could live in a much wider variety of environments (Sandberg 2014:453), and could more efficiently turn energy into conscious life (Bostrom 2003:309).

One might feel sceptical about these scenarios. But given that there are no known scientific obstacles to them, it would be overconfident to be certain, or near-certain, that space

[6] On duration: technological progress brings not only protection against existing extinction risks, but also novel sources of extinction risk (Ord 2020: esp. chs. 4 and 5). On population size: the tendency for richer societies to have lower fertility rates has led some to conjecture that human population, after plateauing around 2100, might significantly decline into the indefinite future, a high 'carrying capacity' notwithstanding (Bricker and Ibbitson 2019).

[7] We return to the likelihood of artificial superintelligence in subsection 4.3.

settlement, or digital sentience, will not occur. Imagine that you could peer into the future, and thereby discovered that Earth-originating civilisation has spread across many solar systems. How surprised would you be, compared to how surprised you would be if you won the lottery?

To move towards particular numbers, we consider three specific future scenarios, taken from Newberry (2021a), where civilisation is: (i) Earthbound; (ii) limited to the Solar System; and (iii) expanded across the Milky Way. In each case, Newberry makes a conservative estimate of the carrying capacity of civilisation in that scenario, on the assumptions that digital life is and is not possible, giving six scenarios in all. He also provides a best-guess estimate of the duration of civilisation in that scenario. These scenarios are not meant to exhaust the space of possibilities, but they give an indication of the potential magnitudes of future population size (see Table 2.1).

To arrive at an *overall* estimate of the expected number of future people, one would further need to estimate probabilities for scenarios such as those in Table 2.1 (and for all other scenarios). However, since the number of lives in the future according to different possible scenarios is spread over many orders of magnitude, in any such expected-value calculation, it tends to be the 'largest' scenario in which one has any non-zero credence that drives the overall estimate. Even a 0.01% credence that biological humanity settles the Milky Way at carrying capacity, for example, contributes at least 10^{32} to the expected number of future beings. Precisely how one's remaining credence is spread among 'smaller' scenarios then makes very little difference.

Because of this, we believe that any reasonable estimate of the *expected* number of future beings is at least 10^{24}. (In fact, we believe that any reasonable estimate must be substantially higher than this; since higher numbers would make little difference to the arguments of this essay, however, we will not press that case here.) However, we are also sympathetic to the concern that if this is the only estimate we consider, the case for strong longtermism would be driven purely by tiny credences in highly speculative scenarios. We will therefore also consider the extent to which the same arguments would go through on some vastly more conservative estimates, as shown in Table 2.2.

Our low estimate (10^{18}) corresponds, for instance, to a 0.0000001% credence in the Solar System (biological life) scenario, with zero credence in either digital sentience or more wide-ranging space settlement. Our restricted estimate (10^{14}) corresponds to the above

Table 2.1 Estimates of the number of future lives in various scenarios.

Scenario	Duration (centuries)	Carrying capacity (lives per century)	Number of future lives
Earth (mammalian reference class)	10^4	10^{10}	10^{14}
Earth (digital life)	10^4	10^{14}	10^{18}
Solar System	10^8	10^{19}	10^{27}
Solar System (digital life)	10^7	10^{23}	10^{30}
Milky Way	10^{11}	10^{25}	10^{36}
Milky Way (digital life)	10^{11}	10^{34}	10^{45}

Table 2.2 Estimates of the expected number of future beings.

	Expected number of future beings
Main estimate	10^{24}
Low estimate	10^{18}
Restricted estimate	10^{14}

estimate for Earthbound life, with zero credence in any larger-population scenario (including both digital sentience and any space settlement). In the arguments that follow, the reader is invited to substitute her own preferred estimate throughout.

We will argue that BR (and hence ASL) holds of society's decision situation even on our restricted estimate, and clearly holds by a large margin on our main estimate.

4 Tractability of significantly affecting the far future

The far-future effects of one's actions are usually harder to predict than their near-future effects. Might it be that the expected instantaneous value differences between available actions decay with time from the point of action, and decay sufficiently fast that in fact the near-future effects tend to be the most important contributor to expected value? If that were so, then neither BR nor ASL would hold.

This is a natural reason to doubt strong longtermism. We will call it *the washing-out hypothesis*.[8]

We agree that the washing-out hypothesis is true of some decision situations. However, we claim that it is false of our society's decision situation.

Given the argument of section 2, our task is to show that there exists at least one option available to society with the property that its far-future expected benefits are significantly greater than the near-future expected benefits of bednet distribution (that is, recall: 0.025 lives saved per $100 spent). We will consider examples in two categories: mitigating extinction risk, and positively shaping the development of artificial superintelligence.

4.1 Influencing the choice among persistent states

Here is an abstract structure which, *insofar as* it is instantiated in the real world, offers a recipe for identifying options whose effects will not wash out.

Consider the space S of all possible fine-grained states the world could be in at a single moment of time (that is, the space of all possible instantaneous microstates). One can picture the history of the universe as a path through this space. Let a *persistent state* be a subset A of S with the property that, given the dynamics of the universe, if the instantaneous state

[8] It is important here to distinguish between *ex ante* and *ex post* versions of the washing-out claim. The *ex post* version is false, as is established by the literature on cluelessness; cf. subsection 7.1. However, it is the much more plausible *ex ante* washing-out claim that is relevant to the arguments of this essay.

of the world is in *A*, then the expected time for which it remains in *A* is extremely long. Now suppose that there are two or more such persistent states, differing significantly from one another in value. Suppose further that the world is not yet in any of the states in question, but might settle into one or the other of the states in question in the foreseeable future. Finally, suppose that there is something we can do now that changes the probability that the world ends up in a better rather than a worse persistent state. Then, as a result of the persistence that is built into the definition, the effects of these actions would not wash out at all quickly.

The empirical question is whether there are, in the real world, any options available that instantiate the structure just described. We claim that there are.

4.2 Mitigating risks of premature human extinction

The non-existence of humanity is a persistent state *par excellence*. To state the obvious: the chances of humanity re-evolving, if we go extinct, are miniscule. Only slightly more subtly, the existence of humanity is also a persistent state: while we face significant risks of premature extinction, as argued in section 3, humanity's *expected* persistence is vast.

These persistent states have unequal expected value. Assuming that on average people have lives of significantly positive welfare,[9] according to total utilitarianism the existence of humanity is significantly better than its non-existence, at any given time. Combining this with the fact that both states are persistent, premature human extinction would be astronomically bad. Correspondingly, even an extremely small reduction in extinction risk would have very high expected value (Bostrom 2013:18). For example, even if there are 'only' 10^{14} lives to come (as on our restricted estimate), a reduction in near-term risk of extinction by 1 millionth of 1 percentage point would be equivalent in value to a million lives saved; on our main estimate of 10^{24} expected future lives, this becomes 10 quadrillion (10^{16}) lives saved.

As is increasingly recognised, as an empirical matter of fact, there are things we could do that would reduce the chance of premature human extinction by a non-negligible amount. As a result, although precise estimates of the relevant numbers are difficult, the far-future benefits of some such interventions seem to compare very favourably, by total utilitarian lights, to the highest available near-future benefits.

The detection and potential deflection of asteroids provides a relatively robust example of such an intervention. This involves scanning the skies to identify asteroids that could potentially collide with Earth and, if one were found, investing the resources to try to deflect it, and/or to prepare bunkers and food stockpiles to help us survive an impact winter. Most of the expected costs here are in detection, because the costs of deflection and preparation are only paid in the very unlikely event that one does detect an asteroid on a collision course with Earth.

In 1996, NASA commenced the Spaceguard Survey, a multi-decade plan to track near-Earth objects with the aim of identifying any on impact trajectories. At a total cost of $71 million by 2012, the Spaceguard Survey had tracked over 90% of asteroids of diameter 1km or more in near-Earth orbit, and all asteroids of diameter 10km or more over 99% of the sky.

[9] We return to this assumption in section 6.

It is not certain that a large asteroid collision would cause human beings to go extinct. We assume a status quo risk of human extinction, conditional on the impact of an asteroid of diameter 10km or more, of 1%. It is also far from certain that we could deflect such a large asteroid, even if we knew it was on a collision course. However, it is far from certain that we could not, and, as above, there are other actions we could take to protect against the extinction risk. We assume here that if such an object were detected to be on a collision course, our deflection and preparation efforts would reduce extinction risk by a proportional 5%. The assumptions in this paragraph follow Newberry (2021b), and seem fairly conservative.

Putting these numbers together, we estimate that the Spaceguard Survey, on average, reduced extinction risk by at least 5×10^{-16} per $100 spent. On our main estimate of the expected number of future beings, this amounts to an additional 500 million lives; this decreases to 500 or 0.05 lives on our low and restricted estimates, respectively.

Of course, we should expect *further* work on asteroids to have lower cost-effectiveness, because of diminishing marginal returns. However, the opportunity remains significant. The remaining risk of a collision with an asteroid with a diameter of 10km or more in the next 100 years has been estimated at 1 in 150 million (Ord 2020:71). It has been estimated that the cost to detect with near-certainty any remaining asteroids of greater than 10km diameter would be at most a further $1.2 billion (Newberry 2021b). On our main (resp., low, restricted) estimate of the expected number of future beings, every $100 of this would, on average, result in 300,000 (resp., 0.3, 0.00003) additional lives. This example therefore supports strong longtermism on our main and low estimates, though not on the restricted estimate. Organisations whose work mitigates risk of extinction from asteroid impacts, and which would benefit from substantially more funding, include the Planetary Society and the B612 Foundation.

While asteroid defence is among the more easily quantified areas of extinction risk reduction, it is far from the only one, or the most significant (Ord 2020: ch. 3). Another possibility concerns global pandemics. Such a pandemic could be natural or man-made, with the latter being particularly concerning (Posner 2004:75–84; Rees 2018: sec. 2.1; Ord 2020). In particular, progress in synthetic biology is very rapid (Meng and Ellis 2020), and it is likely that we will soon be able to design man-made viruses with very high contagiousness and lethality. If such pathogens were released (whether deliberately or by accident (Shulman 2020; Ord 2020:129–131)) in the course of military tensions, or by a terrorist group, there is a real possibility that they could kill a sufficient number of people that the human species would not recover.

In a recent paper, Millet and Snyder-Beattie (2017) use three distinct methods to generate estimates of the risk of an extinction-level pandemic in the next 100 years. The resulting estimates range from 1 in 600,000 to 1 in 50. The authors further use figures from the World Bank to generate a very conservative estimate that $250 billion of spending on strengthening healthcare systems would reduce the chance of such extinction-level pandemics this coming century by at least a proportional 1%.[10]

Taking the geometric mean to average across the two methods that generate the lower estimates for extinction risk, we obtain a risk of about 1 in 22,000 of extinction from a

[10] Two ways in which Millet and Snyder-Beattie's estimate is particularly conservative are (i) that the $250 billion figure is at the extreme upper end of anticipated costs for the intervention they discuss, and (ii) that the intervention in question concerns an extremely broad-based approach to biosecurity, not specifically optimising for extinction risk reduction.

pandemic over the next 100 years.[11] If we use the above figure of $250 billion to reduce the risk by 1%, and assume that the risk reduction occurs throughout the next 100 years but only in that time period, then each $100 of such spending would, in expectation, increase the number of future beings by 200 million (resp., 200, 0.02) on our main (resp., low, restricted) estimate. According to these calculations, the far-future benefits would thereby significantly exceed the near-future benefits of bednet distribution on our main and low estimates of the size of the future, though not on our restricted estimate. Organisations working on these threats include the John Hopkins Center for Health Security, the Nuclear Threat Initiative's biosecurity programme, and Gryphon Scientific.

4.3 Influencing the choice among non-extinction persistent states

A second way of positively impacting the long run is by improving the value of the future conditional on the existence of a very large number of future sentient beings. For concreteness, we focus on one way of doing this: positively shaping the development of artificial superintelligence (ASI), that is, artificial systems that greatly exceed the cognitive performance of humans in virtually all domains of interest.[12]

The idea that the development of sufficiently advanced artificial intelligence could prove a key turning point in history goes back to the early computer pioneers Alan Turing (1951) and I. J. Good (1966). It has more recently been the subject of wider concern.[13] There are two classes of long-term worry.

The first is from *AI-takeover* scenarios (Bostrom 2014; Russell 2019). This worry is that, once we build a human-level artificial intelligence, it would be able to recursively self-improve, designing ever-better versions of itself, quickly becoming superintelligent. From there, in order to better achieve its aims, it will try to gain resources, and try to prevent threats to its survival. It would therefore be incentivised to take over the world and eliminate or permanently suppress human beings. Because the ASI's capability is so much greater than that of humans, it would probably succeed in these aims.

The second worry is from *entrenchment* scenarios (MacAskill 2022). If an authoritarian country were the first to develop ASI, with a sufficient lead, they could use this technological advantage to achieve world domination. The authoritarian leader could then quash any ideological competition. An AI police force could guarantee that potential rebellions are prevented; an AI army would remove any possibility of a coup. And if the leader wanted his ideology to persist indefinitely, he could pass control of society on to an ASI successor before his death. To this end, he could hard-code the goals of the ASI to match his own, have the ASI learn his goals

[11] We use the geometric rather than the arithmetic mean because the estimates in question are spread across several orders of magnitude; the arithmetic mean effectively defers to the highest estimate on the question of order of magnitude. Using the arithmetic mean would lead to results that are still more favourable to strong longtermism. Similarly, we disregard Millet and Snyder-Beattie's 'Model 1' because, as the authors note, this model is flawed in important respects; including this model would also strengthen the case for strong longtermism.

[12] Other areas one might consider here include affecting the values that the world converges on (Reese 2018), or reducing the risk of a totalitarian world government (Caplan 2008).

[13] Those concerned include leading machine learning researchers such as Stuart Russell (2019) and Shane Legg (2008: sec. 7.3), philosophers such as Nick Bostrom (2014), Eliezer Yudkowsky (2013), Toby Ord (2020:138–152), and Richard Ngo (2020), physicists such as Max Tegmark (2017: ch. 4) and Stephen Hawking (2018: ch. 9), and tech entrepreneurs such as Elon Musk (2014), Sam Altman (2015), and Bill Gates (Statt 2015).

from his speech and behaviour, or even 'mind upload', scanning his brain and having it digitally emulated (Sandberg and Bostrom 2008b; Sandberg 2013).

In either of these scenarios, once power over civilisation is in the hands of an ASI, this could persist as long as civilisation does (Finnveden, Riedel, and Shulman 2022). Different versions of the ASI-controlled futures are therefore persistent states with significantly differing expected value, so that we have another instantiation of the structure outlined in subsection 4.1. In the scenario just mentioned, the ASI's goals were "locked in". The ASI would be immortal and all-powerful, so those goals would determine both the trajectory and value of the future. These aims, inherited from its creator, could be pursued across a vast cosmic civilization for billions of years.

Though extinction risks involve dramatic reductions in the size of the future population, these AI scenarios need not. In the classic statement of the AI-takeover scenario, the ASI goes on to settle the stars in pursuit of its goals (Bostrom 2014:100). Similarly, if an authoritarian leader transferred power to an ASI, they too might want their civilisation to be large, populous, and long-lasting. In particular, for a wide variety of goals (such as building the grandest possible temples, doing research, or, in a toy example Bostrom (2014:123–124) gives to illustrate the general phenomenon of misaligned AI, maximising the number of paperclips), acquiring more resources helps with achievement of these goals, which motivates settling the stars. And, in order to fulfil these goals, a populous workforce would be instrumentally valuable. In expectation, the number of future beings, in these scenarios, is very large.

Now, this workforce might consist almost entirely of AIs. But, as we noted in section 3, there are reasons to think that such beings would have moral status, and therefore how well or poorly their lives went would be of moral concern, relevant to the arguments of this essay. And, at least on the authoritarian-takeover scenarios, the ruler might wish to have a very large number of human followers, too.

There are two strands of work aimed at reducing risks from ASI. First, AI safety research, which aims to ensure that AI systems do what we intend them to do (Amodei et al. 2016). Such work is conducted by organisations such as Berkeley's Center for Human-Compatible AI, the Machine Intelligence Research Institute, and labs within Google DeepMind and OpenAI. Second, policy work, in particular to ensure a cooperative approach between countries and companies: for example, by The Partnership on AI, the Centre for the Governance of AI, and the Center for New American Security.

Despite this work, ASI safety and policy are still extremely neglected. For example, the Open Philanthropy Project is the only major foundation with these issues as a key focus area; it spends under $30 million per year on them (Open Philanthropy 2020).[14] The AI safety teams at OpenAI and DeepMind are small.

There is no hard quantitative evidence to guide cost-effectiveness estimates for AI safety work. Expert judgement, however, tends to put the probability of existential catastrophe from ASI at 1–10%.[15] Given these survey results and the arguments we have canvassed, we

[14] Neglectedness is crucial to the argument of this essay. Would strong longtermism still be true if, for example, 10% of global GDP were already spent on the most valuable long-term-oriented interventions? Even if true, would it still be significantly revisionary compared to a near-termist approach, as we have claimed it is at the current margin? We aren't sure. Our claim here is only that the world today is clearly far below this optimum.

[15] Grace et al. (2018) asked 352 leading AI researchers to give a probability on the size of existential risk arising from the development of 'human-level machine intelligence'; the median estimate was 5%. A survey among participants at a conference on global catastrophic risks similarly found the median estimate to be 5% (Sandberg and Bostrom 2008a). One would expect a selection effect to be at work in surveys of those who have chosen to work on existential risk, but not so (or not strongly) for the survey of AI researchers.

think that even a highly conservative assessment would assign at least a 0.1% chance to an AI-driven catastrophe (as bad as or worse than human extinction) over the coming century. We also estimate that $1 billion of carefully targeted spending would suffice to avoid catastrophic outcomes in (at the very least) 1% of the scenarios where they would otherwise occur. On these estimates, $1 billion of spending would provide at least a 0.001% absolute reduction in existential risk. That would mean that every $100 spent had, on average, an impact as valuable as saving 1 trillion (resp., 1 million, 100) lives on our main (resp., low, restricted) estimate—far more than the near-future benefits of bednet distribution. Readers are invited to substitute their own preferred estimate here; even quite large downward adjustments of the figure that we ourselves guess will leave the argument substantially unaffected.

4.4 Uncertainty and 'meta' options

There is a lot of uncertainty in the numbers we have given, even in the most scientifically robust case of asteroid detection. We will give this issue a more thorough treatment in the next section, arguing against various ways in which one might worry it undermines our argument.

One thing that uncertainty can support, however, is a preference for different types of strategy to improve the far future. Rather than directly trying to influence the far future, one could instead try to invest in decision-relevant research, or invest one's resources for use at a later date.

The possibility of either of these strategies strengthens our argument considerably. To see this, let us suppose, for the sake of argument, that no 'first-order' intervention (such as those we discussed in subsections 4.2–4.3) delivers higher far-future expected benefits than the highest available near-future expected benefits, relative to the credences that are appropriate in the present state of information. Suppose, however, that it is highly likely that, conditional on sufficient additional information, at least one of the proposed interventions, or another such intervention (not yet considered) in a similar spirit, would have much higher far-future benefits, relative to the updated credences, than the best available near-future benefits. Then society might *fund research into* the cost-effectiveness of various possible attempts to influence the far future. Provided that subsequent governments or philanthropists would take due note of the results, this 'meta-option' could easily have much greater far-future expected benefits than the best available near-future expected benefits, since it could dramatically increase the expected effectiveness of future governmental and philanthropic action (all relative to *currently* appropriate credences).

A complementary possibility is that, rather than spending now, society could save its money for a later time (Christiano 2014; MacAskill 2019; Trammell 2020). That is, it could set up a sovereign wealth fund, with a longtermist mission. This fund would pay out whenever there becomes available some action that will sufficiently benefit the far future (in expectation), whether that is during the lifetimes of current citizens or later. There would be some annual risk of future governments being misaligned and using the money poorly, but this risk could be mitigated via constitutional enshrinement of the mission, and

would be compensated by the fact that the fund would benefit from compound returns of investment.[16]

These considerations show that the bar that 'intractability' objections to our argument must meet is very high. For BR to fail to hold on such grounds, *every* option available to society must have negligible effect on the far future's expected value. Moreover, it must be near-certain that there will be no such actions in the future, and that no such actions could be discovered through further research. This constellation of conditions seems unlikely.

5 Strong longtermism about individual decisions

So far we have discussed what is best for a society to do, sometimes referring to what billions of dollars would be able to achieve. But what about individuals?

We believe our arguments apply to individuals in much the same way they apply to society as a whole. Suppose Shivani is an individual philanthropist, deciding where to spend her money. Naively, we might think of Shivani as making a contribution to asteroid detection, pandemic preparedness, or AI safety that is proportional to her resources. If $1 billion can decrease the chance of an asteroid collision this century by 1 in 120,000, then $10,000 can decrease the chance of an asteroid collision by 1 in 12 billion. Because the individual's ability to contribute to short-term good would also decrease proportionally, perhaps the argument goes through in just the same way.

This 'naive' argument is, in our view, approximately correct. We foresee three ways of resisting it.

First, one could claim that private individuals are much more limited in their options, to such an extent that Shivani can do nothing to decrease risks from asteroids, pandemics, or AI. However, this is simply not true. Multiple organisations working on these risks, including most of those we mentioned above, accept funding at all scales from private individuals, and would scale up their activity in response.

Second, one could claim that there are increasing returns to scale, so that the impact of a small donation is much less than the relevant fraction of the impact of a large donation. This is an open possibility, but it seems significantly more likely that there are fairly

[16] Plausibly, the gains from the investment would outweigh the risk of value-drift of the fund: the historical real rate of return on risky investments (such as stocks and housing) was around 7% during the period 1870–2015 (Jordà et al. 2019:1228). It seems reasonable to expect substantially lower returns in the future; but even if so, they would still be significantly higher than the risk of future governments misusing the funds; even a 90% probability of a future government misusing the funds over the next century would amount to only 2% annual risk.

There is some precedent for successful long-lasting trusts in the charitable sector. In the US, the John Clarke Trust was founded in 1676 (Ochs, 2019); in the UK, King's School, Canterbury was established in 597 (House of Commons Public Administration Select Committee, 2013). In 1790, Benjamin Franklin invested £1,000 for each of the cities of Boston and Philadelphia: three-quarters of the funds would be paid out after 100 years, and the remainder after 200 years. By 1990, the donation had grown to almost $5 million for Boston and $2.3 million for Philadelphia (Isaacson 2003:473–474). The oldest similar government funds date back to the mid-19th century: Texas's Permanent School Fund was founded in 1845 (Texas Education Agency 2020), and its Permanent University Fund was founded in 1876 (University of Texas System 2021). If the annual chance of failure of such funds were as high as 2%, then the chance of the Texas Permanent School Fund persisting until the present day would be 1 in 30, and the chance of the King's School persisting until the present day would be 1 in 10 trillion. This does not merely appear to be a selection effect: to our knowledge, it is not the case that there have been very large numbers of attempted long-lasting government funds that have failed. This suggests that 2% is a conservatively high estimate of the annual risk of failure.

strongly *diminishing* returns, here as elsewhere.[17] This is for both theoretical and empirical reasons. Theoretically: since interventions vary in their *ex ante* cost-effectiveness, a rational altruistic actor will fund the most cost-effective intervention first, before moving to the next-most cost-effective intervention, and so on. Empirically, diminishing returns have been observed across many fields (e.g. Cassman, Dobermann, and Walters 2002:134; Arnold et al. 2018; Bloom et al. 2020).

Third, one could claim that, once we consider the actions of individuals with smaller amounts of resources, the probability of success from directing those resources to long-term-oriented interventions becomes so low that expected utility theory gives the wrong recommendations. We discuss this issue in section 8.

What of individual decisions about where to direct one's *labour*, rather than one's money? We believe that much the same arguments apply here. Suppose that Adam is a young graduate choosing his career path. Adam can choose to train either as a development economist, or as an AI safety researcher. While there are differences between Adam's decision situation and Shivani's (MacAskill 2014), there are also important similarities. In particular, the considerations that make it better in expectation for Shivani to fund AI safety rather than developing world poverty reduction similarly seem to make it better in expectation for Adam to train as an AI safety researcher rather than as a development economist.

6 Robustness of the argument

In our initial presentation of the argument, we have at times assumed expected total utilitarianism, for simplicity. This raises an important question of how wide a class of axiologies will support axiological strong longtermism.

First, what if instead of maximising expected total welfare, the correct axiology is risk averse?[18] This in fact seems to strengthen the case for strong longtermism: the far-future interventions we have discussed are matters of mitigating catastrophic risks, and in general terms, risk aversion strengthens the case for risk mitigation. With only minor modifications, similar remarks apply if, instead of replacing risk neutrality with risk aversion, we replace appeals to utilitarianism in our argument with (*ex post*) prioritarianism.

Second, if the only means of positively influencing the far future were via reducing the risk of extinction, the case for strong longtermism might rely on controversial views in population ethics, such as totalism, on which the absence of a large number of happy future beings makes things much worse. But many axiologies will not agree that premature extinction is extremely bad. In particular, person-affecting approaches to population ethics tend to resist that claim. According to the spirit of a person-affecting approach, premature

[17] Relatedly, it seems that insofar as scale does make a difference, ASL(i) and (ii) are *more* likely to be true of decision situations involving smaller sums of money, not less likely.

Increasing-returns phenomena are discussed by Pierson (2000).

[18] On the standard account, to be risk averse is to have utility be a concave function of total welfare (Pratt 1964:127; O'Donoghue and Somerville 2018:93). Some have argued that the standard account is inadequate (Rabin 2000; Buchak 2013:30). On risk-weighted expected utility theory, risk aversion is represented by a risk function that transforms the expected utility function (Quiggin 1982; Quiggin and Wakker 1994; Buchak 2013). The differences between these accounts are unimportant for our purposes.

extinction is in itself at worst neutral: if humanity goes prematurely extinct, then there does not exist any person who is worse off as a result of that extinction, and, according to a person-affecting principle, it follows that the resulting state of affairs is not worse. The far-future benefits of extinction risk mitigation may therefore beat the best near-future benefits only conditional on controversial population axiologies.[19]

However, risks from ASI are unlike extinction in this respect: there will be a large population in the future either way, and we are simply affecting how good or bad those future lives are. The idea that it's good to improve expected future well-being *conditional on the existence of a large and roughly fixed-size future population* is robust to plausible variations in population-ethical assumptions.[20]

Third, the example of ASI risk also ensures that our argument goes through even if, in expectation, the continuation of civilisation into the future would be bad (Arrhenius and Bykvist 1995: ch. 3; Benatar 2006; Althaus and Gloor 2018). If this were true, then reducing the risk of human extinction would no longer be a good thing, in expectation. But in the AI lock-in scenarios we have considered, there will be a long-lasting civilisation either way. By working on AI safety and policy, we aim to make the trajectory of that civilisation better, whether or not it starts out already 'better than nothing'.

One feature of expected utilitarianism that *is* near-essential to our argument is a zero rate of pure time preference. With even a modest positive rate of pure time preference (as e.g. on 'discounted utilitarian' axiologies), the argument would not go through. Our assumption of a zero rate, however, matches a consensus that is almost universal among moral philosophers, and also reasonably widespread among economists.[21]

This is of course nowhere near an exhaustive list of possible deviations from expected total utilitarianism. We consider some other deviations below, in the course of discussing cluelessness and fanaticism. Our conclusion is that the case for strong longtermism is at least *fairly* robust to variations in plausible axiological assumptions; we leave the investigation of other possible variations for future research.

[19] It is not immediately clear precisely what a person-affecting approach will say about the value of extinction risk mitigation, since the usual formulations of those theories do not specify how the theories deal with risk, and it is not immediately clear how to extend them to cases that do involve risk. Thomas (2019) explores a number of possibilities.

[20] 'Narrow' person-affecting approaches disagree, since they regard two states of affairs as incomparable whenever those states of affairs have non-identical populations (Heyd 1988). However, such approaches are implausible, for precisely this reason. Similarly, theories on which any two states of affairs with *non-equinumerous* populations are incomparable (Bader n.d.) are implausible. When comparing different sized populations, a 'wide' person-affecting approach will typically map the smaller population to a subset of the larger population, and compare well-being person by person according to that mapping (Meacham 2012); these theories will tend to agree with total utilitarianism on the evaluation of the AI catastrophes we discuss.
For similar reasons, we also do not consider here the incomparability that is introduced by a 'critical range' view (Blackorby, Bossert, and Donaldson 1996).

[21] A zero rate of pure time preference is endorsed by, inter alia, Sidgwick (1890), Ramsey (1928), Pigou (1932), Harrod (1948), Solow (1974), Cline (1992), Cowen (1992), Stern (2007), Broome (2008), Dasgupta (2008), Dietz, Hepburn, and Stern (2008), Buchholz and Schumacher (2010), and Gollier (2013). In a recent survey of academic economists with expertise on the topic of social discounting, 38% of respondents agreed with this 'Ramsey-Stern view' (Drupp et al. 2018:119). Greaves (2017) provides a survey of the arguments on both sides.

Even among philosophers, the consensus against discounting future well-being is not universal. In particular, some plausible models of partiality suggest assigning greater effective moral weight to one's own contemporaries than to far-future people (Setiya 2014; Mogensen 2019). However, even these models seem unlikely to recommend *sufficient* discounting to undermine the argument for longtermism (Mogensen 2019: sec. 6).

7 Cluelessness

Section 4 focused on worries about our *abilities to affect* the far future. A distinct family of worries is more directly epistemic, and involves the idea that we are *clueless* both about what the far future will be like, and about the differences that we might be able to make to that future.[22] Perhaps the beings that are around will be very unlike humans. Perhaps their societies, if they have anything that can be called a society at all, will be organised in enormously different ways. For these and other reasons, perhaps the kinds of things that are conducive to the well-being of far-future creatures are very different from the kinds of things that are conducive to our well-being. Given all of this, can we really have any clue about the far-future value of our actions *even in expectation*?

We take it for granted that we cannot *know* what the far future will be like. But, since the argument of sections 2–6 has already been conducted in terms of *expected* value, lack of knowledge cannot ground any objection to the argument. The objection must instead be something else.

In fact, there are several quite distinct possibilities in the vicinity of the 'cluelessness' worry. In the present section, we address five of these objections, relating to simple cluelessness, conscious unawareness, imprecision, arbitrariness, and ambiguity aversion.

7.1 Simple cluelessness

Our concern is with relatively weighty decisions, such as how to direct significant philanthropic funding. But it is illuminating to compare these to far more trivial decision situations, such as a choice of whether or where to go shopping on a given day.

Even in the latter cases, many have argued, we can be all but certain that our choice will have highly significant consequences *ex post*—far more significant than the more predictable nearer-term effects.[23] The reasons for this include the tendency for even trivial actions to affect the identities of future persons far into the future. However, when comparing quite trivial alternatives, we can have no idea *which* of the two will turn out to be superior vis-à-vis these deeply unpredictable very far-future effects.

Some have argued that these facts undermine any attempt to base decisions on considerations of the overall good *even in trivial everyday decision contexts* (e.g. Lenman 2000). We agree with Greaves (2016) that this concern is overblown: in the context of relatively trivial everyday decisions, at least, the deeply unpredictable far-future effects plausibly cancel out for the purpose of comparing actions in expected-value terms. Consequently, there is no objection here to basing *these* decisions on an expected-value assessment of nearer-future, more foreseeable effects.

As we have argued in section 4, however, decisions about how to spend philanthropic funding are disanalogous in this respect. We are not discussing the possibility that either funding AI safety research or not funding it might lead, as chance has it, to the birth of

[22] Since 'washing-out' concerns whether we are able to affect the far future *in expectation*, this too has an epistemic aspect, so that the distinction between the concerns of section 4 and those discussed here is not completely clear (Tarsney 2019). Nonetheless, the issues raised seem sufficiently different to warrant a separate treatment.

[23] See e.g. Lenman (2000), Greaves (2016). We agree with this claim, but our argument does not rely on it.

an additional unusually good or bad person several centuries hence. Rather, we are discussing the possibility that funding AI safety might have its intended effect of making AI safer. While there are certainly severe uncertainties in such work, it would be overly pessimistic to insist that success is no more likely than counterproductivity. Considerations of such 'simple' cluelessness therefore do nothing to undermine the argument for strong longtermism.

7.2 Conscious unawareness

The expected-value approach we assumed in section 3 is intended as a *subjective* decision theory: that is, it utilises only material that is accessible to the decision-maker at the time of decision. In particular, therefore, there is an implicit assumption that the agent herself is in a position to grasp the states, acts, and consequences that are involved in modelling her decision.

But perhaps this is not true. Consider, for example, would-be longtermists in the Middle Ages. It is plausible that the considerations most relevant to their decision—such as the benefits of science, and therefore the enormous value of efforts to help make the scientific and industrial revolutions happen sooner—would not have been on their radar. Rather, they might instead have backed attempts to spread Christianity, perhaps by violence: a putative route to value that, by our more enlightened lights today, looks wildly off the mark.

The suggestion, then, is that our current predicament is relevantly similar to that of our medieval would-be longtermists. Perhaps there are actions available to us that would, *if* we were able to think it all through in full detail, then deliver high expected benefits for the far future. But we know, if only by induction from history, that we have not thought things through in all relevant detail. Perhaps we thereby have good reason to reject subjective expected-value analysis, and use some quite different form of decision analysis to assess far-future effects—in which case, all bets are as yet off regarding what the conclusion will be.

This is the issue of *conscious unawareness*—knowing that one is unaware of many relevant considerations, mere awareness of which would influence one's decision-making. Following much of the recent literature on this topic, however, our view is that conscious unawareness does not occasion any particularly significant revision of the Bayesian framework, for three reasons.

First, we know that we operate with coarse-grained models, and that the reasons for this include unawareness of some fine-grainings. Of course, failure to consider key fine-grainings might lead to different expected values and hence to different decisions, but this seems precisely analogous to the fact that failure to possess more information about which state in fact obtains similarly affects expected values (and hence decisions). Since our question is which actions are *ex ante* rational, both kinds of failure are beside the point.

Second, we know we are likely to be omitting some important possible states of nature from our model altogether. But consciousness of this can be modelled by inclusion of a 'catchall' state: 'all the other possibilities I haven't thought of'. Again, conceptualising parts of this state in more explicit terms might change some expected-value assessments, but

again this does nothing to undermine the *ex ante* rationality of decisions taken on the basis of one's existing assessments.[24]

Third, while the best options might well be ones that have not occurred to us, that does nothing to impugn the rationality of assessments of those possible options that *have* occurred to us. And our argument for strong longtermism, recall, requires only a lower bound on attainable far-future expected benefits.

We do not claim (nor do we believe) that issues of conscious unawareness have no effect on what the reasonable credences and values in a given decision situation are. The point is rather that these issues need not occasion any deep structural change to the analysis. Our further claim is that the numbers we have suggested in section 4 are reasonable *after* taking issues of conscious unawareness into account.

7.3 Arbitrariness

An obvious and potentially troubling feature of our discussion in section 4 is the paucity of objective guidance for the key values and probabilities. This seems to contrast starkly with, for instance, the usual impact evaluations for the short-term benefits of bednet distribution, which can be guided by relatively hard evidence (GiveWell 2020b).

This gives rise to three distinct, though related, concerns with the standard Bayesian approach that we have used. The first is simply that the probabilities and/or values in this case are too arbitrary for our argument to carry any weight. The second is that in cases where any precise assignments would be this arbitrary, it is inappropriate to have precise credences and values at all. The third is that in such cases, the appropriate decision theory is ambiguity averse, and that this might undermine the argument for strong longtermism. We address these concerns in turn.

The 'arbitrariness' objection is that even if a rational agent must have some precise credence and value functions, there is so little by way of rational restriction on *which* precise functions are permissible that the argument for strong longtermism is little more than an assertion that the authors' own subjective probabilities are ones relative to which this thesis is true.

We have some sympathy with this objection. However, there is a distinction between there being no watertight argument against some credence function on the one hand, and that credence function being *reasonable* on the other. Even in the present state of information, in our view credence-value pairs such that the argument for strong longtermism fails are unreasonable. If, for instance, one had credences such that the expected number of future people was only 10^{14}, the status quo probability of catastrophe from AI was only 0.001%, *and* the proportion by which $1 billion of careful spending would reduce this risk was also only 0.001%, then one would judge spending on AI safety equivalent to saving only 0.001 lives per $100—less than the near-future benefits of bednets. But this constellation of conditions seems unreasonable.

However, we note that this issue is contentious. We regard the quantitative assessment of the crucial far-future-related variables as a particularly important topic for further research.

[24] The first type of unawareness is unawareness of possible *refinements*, the second is unawareness of possible *expansions* (Bradley 2017: sec. 12.3; Stefánsson and Steele 2021: sec. 3.2).

7.4 Imprecision

Imprecise approaches represent an agent by a *class* of pairs of probability and value functions—a *representor*—rather than a single such pair. The natural interpretation is that these correspond to incomplete orderings of options: one option is better than another, for instance, if and only if the first has higher expected value than the second on *all* probability-value pairs in the representor.[25]

ASL involves comparing *ex ante* far-future benefits with *ex ante* near-future or total benefits. If imprecision is a feature of rational evaluation at all, it is plausibly a particularly prominent feature of evaluation of far-future consequences. So perhaps, for any option (including the ones we have discussed above), any reasonable representor contains at least some elements according to which the far-future benefits of this option are no higher than the near-future benefits of bednet mitigation.

It is somewhat complex to say how one should evaluate ASL in the context of such imprecision. (For instance: Should we simply evaluate ASL itself relative to each element of the representor in turn, and supervaluate to arrive at an overall verdict? Or should we seek to define subsentential terms like 'near-best' in the context of representors? If the latter, how exactly?) The general idea, though, is that one way or another, if the possibility in the last sentence of the preceding paragraph is realised, then ASL is at least not determinately true.

Our reply to the imprecision critique is very similar to our reply to the arbitrariness critique. While we do not take a stand on whether or not any imprecision of valuation is either rationally permissible or rationally required (Elga 2010), we don't ourselves think that any plausible degree of imprecision in the case at hand will undermine the argument for strong longtermism. For example, we don't think any reasonable representor even *contains* a probability function according to which efforts to mitigate AI risk save only 0.001 lives per $100 in expectation. This does seem less clear, however, than the claim that this is not a reasonable precise credence function.

7.5 Ambiguity aversion

In employing the standard Bayesian machinery, we have been assuming *ambiguity neutrality*. In contrast, an *ambiguity-averse* decision theory favours gambles that involve more rather than less objectively specified probabilities, other things being equal (Machina and Siniscalchi 2014).

Empirically, people commonly demonstrate ambiguity aversion. Suppose, for example, that one urn contains 50 red balls and 50 black balls, and a second urn contains both red and black balls in unknown proportion (Ellsberg 1961). If one is ambiguity averse, one might strictly prefer to bet on the risky urn, where one knows the probability of winning, regardless of which colour one is betting on. This preference seems inconsistent with expected utility theory, but is widespread (Trautmann and Kuilen 2015).

It might seem at first sight that ambiguity aversion would undermine the case for strong longtermism. In contemplating options like those discussed in section 4, one needs to settle

[25] Bewley (2002), Dubra, Maccheroni, and Ok (2004), and Galaabaatar and Karni (2013) provide representation theorems linking such representations to incomplete orderings.

one's credence that some given intervention to reduce extinction risk, or to increase the safety of ASI, would lead to a large positive payoff in the far future. But again, there seems significant arbitrariness here. In contrast, impact evaluations for the near-future benefits of bednet distribution seem to involve much more precisely bounded probabilities. Might an ambiguity-averse decision theory, then, take a substantially dimmer view of the far-future benefits of existential risk mitigation, and hence of strong longtermism?

Our answer is 'no', for two reasons.

First, whether or not ambiguity aversion has any prospect of undermining the argument for strong longtermism depends, in the first instance, on whether the agent in question is ambiguity averse with respect to the state of the world, or instead with respect to the difference one makes oneself to that state. The above argument-sketch implicitly assumed the latter.[26] But, if one is going to be ambiguity averse at all, it seems more appropriate for an altruist to be ambiguity averse in the former sense (Greaves et al, 2024). And it is far from clear that actions seeking to improve the far future increase ambiguity with respect to the state of the world. It is already extremely ambiguous, for instance, how much near-term extinction risk humanity faces.[27] We see no reason to think that this latter ambiguity is increased, rather than decreasing or remaining the same, by, for example, funding pandemic preparedness.[28]

Second, although it is psychologically natural, and correspondingly widespread, ambiguity aversion is anyway irrational. Here we agree with a fairly widespread consensus; we have nothing to add to the existing debate on this question.[29]

We conclude that the possibility of ambiguity aversion does not undermine the argument for strong longtermism.

8 Fanaticism

One obvious point of contrast between the paradigm examples of ways to attain high near-future vs. far-future expected benefits is that the former tend to involve high probabilities of relatively modest benefits, whereas the latter tend to involve tiny probabilities of enormous benefits. In discussing actions aimed at mitigating extinction risk, for instance, we conceded that it is *very unlikely* that any such action makes any significant difference; the argument for prioritising such actions nonetheless is characteristically that *if they do* make a significant difference, they might make a truly enormous one.

Even among those who are sympathetic in general to expected utility theory, many balk at its apparent implications for cases of this latter type. Suppose you are choosing between a 'safe option' of saving a thousand lives for sure and a 'risky option' that gives a 1 in 1

[26] To see the distinction in Ellsberg's two-urns setting, suppose that in the status quo, one is set to receive $100 iff the ambiguous urn delivers a red ball. Suppose one's choice is between whether to *add to that background gamble* a bet on a black ball being drawn from the risky urn, or instead from the ambiguous urn. Pretty clearly, ambiguity aversion in the standard sense will recommend the latter (since one then faces zero ambiguity overall), notwithstanding the fact that the benefit *delivered by one's action* is more ambiguous in this case.

[27] Beard, Row, and Fox (2020, Appendix A) and Sandberg and Bostrom (2008a) both present a wide range of estimates from around 1% to 50%, from (respectively) a literature review and a conference participant survey.

[28] We investigate the issues outlined in this paragraph in more depth in (Greaves et al, 2024).

[29] See e.g. (Al-Najjar and Weinstein 2009) for a survey of arguments that ambiguity aversion is irrational. Rowe and Voorhoeve (2018) and Stefánsson and Bradley (2019) defend its rationality.

trillion chance of saving a quintillion lives. The expected number of lives saved is a thousand times greater for the risky option. Unless the utility function is very non-linear as a function of lives saved, correspondingly, the expected utility of the latter option is also likely to be greater. Yet, if you choose the risky gamble, it is overwhelmingly likely that a thousand people will die, for no gain.[30]

Intuitively, it seems at least permissible to save the thousand in this case. If so, this might suggest that while expected utility theory is a good approach to choice under uncertainty in more ordinary cases, it fails in cases involving extremely low probabilities of extremely large values. One might, then, seek a 'non-fanatical' decision theory—one that does not require the agent to sacrifice arbitrarily much, with probability arbitrarily close to 1 in 'fanatical' pursuit of an extremely unlikely but enormously larger payoff. Might a non-fanatical decision theory undermine the case for strong longtermism?

We regard this as one of the most plausible ways in which the argument for strong longtermism might fail. Our view is that at present, the question cannot be confidently settled, since research into the possibility of a non-fanatical decision theory is currently embryonic. However, initial results suggest that avoiding fanaticism might come at too high a price.

Beckstead and Thomas (2020), for instance, consider a sequence of gambles. The first gamble delivers a large but relatively modest benefit with certainty. The last gamble delivers an enormously large benefit with extremely small probability, and zero benefit otherwise. These two gambles are linked by a sequence in which each gamble offers only a very *slightly* lower probability of winning than the previous gamble, and involves a *much* better benefit if one does win. This sequence-schema illustrates that any transitive theory that is not fanatical must instead be worryingly 'timid': in at least one pairwise comparison of adjacent gambles, even an arbitrarily large increase in the value of a positive payoff fails to compensate for any arbitrarily small decrease in its probability. As Beckstead and Thomas go on to show, such timidity in turn leads to implausibly extreme forms of risk aversion in some cases, and to particularly implausible forms of dependence of option-assessments on assessments of causally isolated aspects of the state of affairs.

A complementary reply is that in any case, the probabilities involved in the argument for longtermism might not be sufficiently extreme for any plausible degree of resistance to 'fanaticism' to overturn the verdicts of an expected-value approach, at least at the societal level. For example, it would not seem 'fanatical' to take action to reduce a 1 in 1 million risk of dying, as one incurs from cycling 35 miles or driving 500 miles (respectively, by wearing a helmet or wearing a seat belt (Department of Transport 2020)). But it seems that society can positively affect the very long-term future with probabilities well above this threshold. For instance, in subsection 4.3, we suggested a lower bound of 1 in 100,000 on a plausible credence that $1 billion of carefully targeted spending would avert an existential catastrophe from artificial intelligence.

Things are less clear on the individual level. If, for example, $10 billion can reduce the risk of extinction (or a comparably bad outcome) by 1 in 100,000, and an individual philanthropist makes a $10,000 contribution with effects proportional to that, then the philanthropist would reduce extinction risk by 1 in 10 billion. At this level, we are unlikely to find commonplace decisions relying on that probability that we would regard as

[30] A similar example is that of Pascal's Mugging (Bostrom 2009).

non-fanatical.[31] So, if one is inclined to take seriously the fanaticism worry, despite the problems with 'timidity', it may be that the probabilities in question are problematically small on the individual level, but not at the social level.

Our inclination is to think that our intuitions on the societal level are correct, and that our intuitions around how to handle very low probabilities are unreliable. The latter has some support from the psychological literature (Kahneman and Tversky 1979:282–283; Erev, Glozman, and Hertwig 2008).

We therefore tentatively conclude that considerations of fanaticism do not undermine the argument for strong longtermism.

9 Deontic strong longtermism

In section 2, we distinguished between axiological and deontic versions of strong longtermism. So far, our discussion has focused exclusively on the case for the axiological claim.

The deontic analogue to ASL is

> **Deontic strong longtermism (DSL):** In the most important decision situations facing agents today,
>
> (i) One ought to choose an option that is near-best for the far future.
>
> (ii) One ought to choose an option that delivers much larger benefits in the far future than in the near future.

Just as ASL concerns *ex ante* axiology, the 'ought' in DSL is the subjective ought: the one that is most relevant for action-guidance, and is relative to the credences that the decision-maker ought to have.[32]

Without assuming consequentialism, DSL does not immediately follow from ASL. We believe, however, that our argument for ASL naturally grounds a corresponding argument for DSL. This is because of the following *stakes-sensitivity argument*:

> (P1) When the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor, one ought to choose a near-best option.
>
> (P2) In the most important decision situations facing agents today, the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor.
>
> (C) So, in the most important decision situations facing agents today, one ought to choose a near-best option.

[31] One exception might be putting on a seatbelt for a 1-mile drive. If doing so decreases one's chance of a fatal accident by a factor of one-third, then the seatbelt reduces one's risk of death by about 1 in 1 billion. But perhaps this is not our reason for wearing seatbelts for short journeys.

[32] It is widely agreed that either it is useful to distinguish between objective and subjective senses of 'ought' (Ewing 1948:118–122; Brandt 1959:360–367; Russell 1966; Parfit 1984:25; Gibbard 2005; Parfit 2011; Portmore 2011; Dorsey 2012; Olsen 2017), or 'ought' is univocal and subjective (Prichard 1932; Ross 1939:139; Howard-Snyder 2005; Zimmerman 2006; 2008; Mason 2013). Our discussion presupposes that one of these disjuncts is correct. A minority of authors holds that 'ought' is univocal and objective (Moore 1903:199–200, 229–230; Ross 1930:32; Thomson 1986:177–179; Graham 2010; Bykvist 2011). Similarly (but less discussed), one might be sceptical of the notion of *ex ante* axiology; again, our discussion of ASL has presupposed that any such scepticism is misguided.

DSL follows from the conjunction of (C) and ASL.

The stakes-sensitivity argument is obviously valid. Are its premises true?

(P1) appeals to only a very moderate form of stakes-sensitive non-consequentialism. It allows that there may be some actions that are always permissible or prohibited, no matter how great the axiological stakes: for example, perhaps one is always permitted to save the life of one's child; or perhaps one is always prohibited from torturing another person. And it only entails that comparatively minor prerogatives are overridden when the stakes are very high.[33]

It is highly plausible that there should be at least this much stakes-sensitivity. The lack of stakes-sensitivity is a common objection to Kant's notorious view that even if a friend's life depends on it, one should not tell a lie (Kant 1996). Turning to prerogatives, in 'emergency situations' like wartime, ordinary prerogatives—for instance, to consume luxuries, to live with one's family, and even to avoid significant risks to one's life—are quite plausibly overridden. Nagel (1978) observes that public morality tends to be more consequentialist in character than private morality; one natural partial explanation for this (though not the one emphasised by Nagel himself) is that in public contexts such as governmental policy decisions, the axiological stakes tend to be higher.

We foresee two lines of resistance to (P1). First, one could reject the idea of 'the good' altogether (Thomson 2008: sec. 1.4). On this view, there is simply no such thing as axiology. It's clear that our argument as stated would not be relevant to those who hold such views. But such a view must still be able to explain the fact that, in cases where there is a huge amount at stake, comparatively minor constraints and prerogatives get overridden. It seems likely that any such explanation will result in similar conclusions to those we have drawn, via similar arguments.

Second, and more plausibly, perhaps only some sorts of axiological considerations are relevant to determining what we ought to do. We consider two ways in which this idea might undermine our argument.

First, on a non-aggregationist view, comparatively small *ex ante* benefits to individuals are not relevant to determining what one ought to do, even if the benefits apply to an enormous number of people (Scanlon 1998:235; Voorhoeve 2014; Frick 2015).

Second, perhaps axiological considerations cannot outweigh non-consequentialist considerations when the axiological considerations involve altering the identities of who comes into existence (Parfit 1984: ch. 16).

However, both lines of thought risk proving too much. Let's first consider the non-aggregationist response. Consider a Briton, during WWII, deciding whether to fight; or someone debating whether to vote in their country's general election; or someone deciding whether to join an important political protest; or someone deciding whether to reduce their carbon footprint. In each case, the *ex ante* benefits to any particular other person are tiny. But in at least some such cases, it's clear that the agent is required to undertake the relevant action, and the most natural explanation of why is because the axiological stakes are so high.[34]

Second, consider the non-identity response. It's clear that governments ought to take significant action to fight climate change. But almost all of the expected damages from climate

[33] (P1) is very similar to Singer's claim that 'If it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it' (Singer 1972:231).

[34] None of these examples, however, involves foregoing an opportunity to save many lives of identified people. In this respect, our examples are perhaps relevantly dissimilar to a decision between spending to benefit the far vs. the near future. We thank an anonymous referee for pressing this reply.

change come from its impacts on those who are yet to be born.[35] What's more, any policy designed to mitigate climate change will also affect the identities of those unborn people. Endorsing the non-identity response would therefore risk rejecting the idea that welfarist considerations generate any obligations for society today to fight climate change, even while accepting that climate change will significantly and avoidably reduce welfare in expectation for centuries to come. That position is clearly incorrect.

Turning now to (P2): The 'high-stakes' aspect of this premise is justified in part on the basis of the arguments of sections 3–4. At least on our main and low estimates of the expected size of the future, in the decision situations we've discussed, not only are the best options those that have the near-best far-future consequences, but they are *much* better than those options whose far-future consequences are nowhere near best.

At the same time, at least for most members of rich countries, the decision situations we've discussed are those where the personal prerogatives are arguably comparatively minor, and where there are no serious side-constraints. This is clearest in the cases of individual decisions about *where to direct* one's altruistic spending (holding fixed the total size of one's 'altruistic budget'), and about career choice. The decision to give to organisations that will positively influence the far future rather than organisations more geared towards improving the near future, or to work in a career that is particularly beneficial for the long-term future, might well involve some sacrifices.[36] But they are not close to the sorts of sacrifices where there might be absolute or near-absolute prerogatives. Similarly, these are not circumstances where one is required to violate side-constraints in order to achieve the near-best long-term outcome.

The slightly less clear cases are those involving individual decisions about the total *size* of one's 'altruistic budget' (vs. 'personal budget'), and societal decisions about how many resources to devote to improving the prospects for the far future (vs. the near future, including the lifetimes of present people). Here, it remains true that no serious side-constraints need be involved. One might worry, though, that here our argument will be too demanding: might it imply that we, individually or as a society, ought to devote *most* of our resources to improving the far future, at the large expense of our own prudential interests?

As in the discussion of demandingness in the context of global poverty, a range of responses to this concern is possible. We have nothing to add to the existing literature on demandingness (e.g. Kagan 1984; Mulgan 2001; Hooker 2009).[37] We will simply note that even if, for example, there is an absolute cap on the total sacrifice that can be morally required, it seems implausible that society today is currently anywhere near that cap. The same remark applies to at least the vast majority of individuals in rich countries. We ought to be doing a lot more for the far future than we currently are.[38]

[35] For example, the Stern Review predicts the vast majority of damages to occur after 2100 in both 'baseline' and 'high climate' scenarios (Stern 2007:178, fig. 6.5d).

[36] There are, however, reasons to think that these sacrifices are not as great as we might initially suppose (MacAskill, Mogensen, and Ord 2018).

[37] Mogensen (2021) discusses specifically the relationship between demandingness and longtermism.

[38] Might our arguments go further than this, and justify atrocities in the name of the long-term good? Perhaps the French Revolution had good long-term consequences, in terms of bringing about a more liberal and democratic world: does strong longtermism, if so, justify the guillotine?

We do not think so, for at least two reasons. The first is that, for such serious side-constraints, something closer to absolutism or near-absolutism becomes much more plausible (or, at least, it takes more than mere *ex ante* goodness to justify violation of those side-constraints). The second is that, in almost all cases, when there is some option available that promotes the long-term good while violating a serious side-constraint, there will be some alternative option available that achieves a similar amount of long-term good without violating that side-constraint. Liberal democracy could have been achieved in France without the Reign of Terror.

10 Summary and conclusions

The potential future of civilisation is vast. Once we appreciate this, it becomes plausible that impact on the far future is the most important feature of our actions today.

Strong longtermism would be false in a world that had sufficiently weak causal connections between the near and the distant future, such that it was too difficult to significantly influence the course of the very long-run future. However, we have argued, the world we find ourselves in today does not have this feature.

We presented our central case in terms of (i) a total utilitarian axiology and (ii) an expected-value treatment of decision-making under uncertainty. However, we argued, plausible deviations from either or both of these assumptions do not undermine the core argument.

This essay mainly focused on the decision situations of a society or individual considering how to spend money without constraints as to cause area, and of an individual's career choice. We argued that these are situations where we can in expectation significantly influence the far future. Precisely because of this, they are among the most important decision situations we face, and axiological strong longtermism follows.

In our own view, the weakest points in the case for axiological strong longtermism are the assessment of numbers for the cost-effectiveness of particular attempts to benefit the far future, the appropriate treatment of cluelessness, and the question of whether an expected-value approach to uncertainty is too 'fanatical' in this context. These issues in particular would benefit from further research.

In addition to axiological issues, we also discussed the counterpart deontic issues. We suggested that deontic strong longtermism might well be true even if consequentialism is false, on the grounds that (i) the stakes involved are very high, (ii) a plausible non-consequentialist theory has to be sensitive to the axiological stakes, becoming more consequentialist in output as the axiological stakes get higher, and (iii) in the key decision situations, any countervailing constraints and/or prerogatives are comparatively minor. Quite plausibly, in the world as it is today, the most important determinants of what we ought to do arise from our opportunities to affect the far future.

It is possible, but far from obvious, that far-future impacts are also more important than near-future impacts in a much wider class of decision situations: for instance, decisions about whether or not to have a child, and government policy decisions within a relatively narrow 'cause area'. Insofar as they are, strong longtermism could potentially set a methodology for further work in applied ethics and applied political philosophy: for each issue in these subfields, one could identify the potential far-future effects from different actions or policies, and then work through how these bear on the issue in question. The answers might sometimes be surprisingly revisionary.

Appendix

We claimed in the main text that (BR) entails:

(a) that ASL(i) holds of a restriction of society's decision situation, obtained by removing any options involving net expected short-term harm from the choice set; and
(b) that ASL(ii) holds of society's decision situation.

Here, we make these claims precise, and supply the proofs for them.

Terminology and notation

For any option x, let $N(x)$, $F(x)$, $V(x)$ respectively denote x's near-future, far-future, and overall benefits. Let N^*, F^*, V^* respectively be the highest available near-future, far-future, and overall benefits. Let F' be the highest far-future benefit that is available without net short-term harm.

We interpret both 'near-best overall' and 'near-best for the far future' in terms of proportional distance from zero benefit to the maximum available benefit, and 'much larger' in terms of a multiplicative factor. There is, of course, flexibility on the precise values of the factors involved. We therefore consider the following precisifications of our key claims, carrying free parameters:

BR(n): $F' \geq nN^*$.

$ASL_i(\epsilon_O, \epsilon_F)$: Every option that delivers overall benefits of at least $(1-\epsilon_O)V^*$ delivers far-future benefits of at least $(1-\epsilon_F)F^*$.

$ASL_{ii}(\epsilon_O, r)$: Every option that delivers overall benefits of at least $(1-\epsilon_O)V^*$ delivers far-future benefits that are at least r times its own near-future benefits.

In what follows, we prove claims (a) and (b) for specified relationships between the parameter values.

Precisification of claim (a). We claim (more precisely) that if BR(n) holds of a given decision situation, then for any $\epsilon_O \in [0,1]$, $ASL_i(\epsilon_O, \epsilon_O + \frac{1}{n})$ holds of the restricted decision situation (with any options involving net short-term harm removed). For example, if $n = 10$, then every option that delivers at least 90% of available overall expected benefits delivers at least 80% of available far-future expected benefits, once any options involving net short-term harm are ruled out.

Proof. Suppose that BR(n) holds. Since far-future benefit F' is attainable without near-future net harm, the overall best option must deliver total benefits of at least F'; so any near-best option must deliver total benefits of at least $(1-\epsilon_O)F'$. But by BR(n), the maximum attainable near-future benefit is at most $\frac{F'}{n}$. Therefore, any near-best option must deliver far-future benefits of at least $(1-\epsilon_O - \frac{1}{n})F'$. But in this decision situation, $F' = F^*$ (since near-future net harm is here ruled out).

Precisification of claim (b). We claim (more precisely) that if BR(n) holds then for any $\epsilon_O \in [0,1]$, $ASL_{ii}(\epsilon_O, (1-\epsilon_O)n-1)$ also holds. For example, if $n = 10$, then every option that delivers at least 90% of available overall expected benefits delivers at least eight times as much far-future as near-future expected benefit.

Proof. Let x be any option that is near-best overall. Then

$$
\begin{aligned}
V(x) &\geq (1-\epsilon_O)V^* && \text{by definition of near-best} \\
&\geq (1-\epsilon_O)F' && \text{since, by hypothesis, } F' \text{ is achievable without short-term harm}
\end{aligned}
$$

But $V(x) = N(x) + F(x)$, so it follows that

$$
\begin{aligned}
F(x) &\geq (1-\epsilon_O)F' - N(x) \\
&\geq (1-\epsilon_O)F' - N^* \\
&\geq ((1-\epsilon_O)n-1)N^* && \text{by BR(n)} \\
&\geq ((1-\epsilon_O)n-1)N(x)
\end{aligned}
$$

References

Adams, F. C. (2008), 'Long-Term Astrophysical Processes', in N. Bostrom and M. Cirkovic (eds.), *Global Catastrophic Risks* (Oxford University Press), 33–47.

Adams, F. C. and Laughlin, G. (1999), *The Five Ages of the Universe: Inside the Physics of Eternity* (Free Press).

Al-Najjar, N. I. and Weinstein, J. (2009), 'The Ambiguity Aversion Literature: A Critical Assessment', *Economics & Philosophy 25*/3: 249–284.

Althaus, D. and Gloor, L. (2018), 'Reducing Risks of Astronomical Suffering: A Neglected Priority', *Foundational Research Institute*, https://foundational-research.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/ (accessed 22 January 2025).

Altman, S. (2015), 'Machine Intelligence, Part 1', https://blog.samaltman.com/machine-intelligence-part-1

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016), 'Concrete Problems in AI Safety', https://arxiv.org/abs/1606.06565

Armstrong, S. and Sandberg, A. (2013), 'Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox', *Acta Astronautica* 89:1–13.

Arnold, R., Scott Marcus, J., Petropoulos, G., and Schneider, A. (2018), 'Is Data the New Oil? Diminishing Returns to Scale', *29th European Regional Conference of the International Telecommunications Society*.

Arrhenius, G. and Bykvist, K. (1995), 'Future Generations and Interpersonal Compensations: Moral Aspects of Energy Use', *Uppsala Prints and Preprints in Philosophy* 21 (Uppsala Universitet).

Bader, R. (n.d.), 'Person-Affecting Population Ethics' (unpublished manuscript).

Barnosky, A., Matzke, N., Tomiya, S., Wogan, G., Swartz B., Quental, T., Marshall, C., McGuire, J., Lindsey, E., Maguire, K., Mersey, B., and Ferrer E. (2011), 'Has the Earth's Sixth Mass Extinction Already Arrived?' *Nature* 471/7336: 51–57.

Beard, S., Rowe, T., and Fox J. (2020), 'An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards', *Futures* 115: 102469.

Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future*, PhD thesis, Rutgers University, https://rucore.libraries.rutgers.edu/rutgers-lib/40469/PDF/1/play/

Beckstead, N. (2014), 'Will We Eventually Be Able to Colonize Other Stars? Notes from a Preliminary Review', https://www.fhi.ox.ac.uk/will-we-eventually-be-able-to-colonize-other-stars-notes-from-a-preliminary-review/ (accessed 22 January 2025).

Beckstead, N. and Thomas, T. (2020), 'A Paradox for Tiny Probabilities and Enormous Values', GPI Working Paper 10-2020 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/ (accessed 22 January 2025)

Benatar, D. (2006), *Better Never to Have Been: The Harm of Coming into Existence* (Clarendon Press).

Berger, A. (2019), 'GiveWell's Top Charities Are (Increasingly) Hard to Beat', *Open Philanthropy Project*, https://www.openphilanthropy.org/blog/givewells-top-charities-are-increasingly-hard-beat (accessed 26 January 2021).

Bewley, T. F. (2002), 'Knightian Decision Theory. Part I', *Decisions in Economics and Finance*, *25*/2: 79–110.

Blackorby, C., Bossert, W., and Donaldson, D. (1996), 'Quasi-orderings and Population Ethics', *Social Choice and Welfare* 13: 129–150.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020), 'Are Ideas Getting Harder to Find?', *American Economic Review* 110/4: 1104–1144.

Bollard, L. (2016), 'Initial Grants to Support Corporate Cage-free Reforms', *Open Philanthropy Project*, https://www.openphilanthropy.org/blog/initial-grants-support-corporate-cage-free-reforms#Corporate_cage-free_campaigns_are_extremely_cost-effective (accessed 26 January 2021).

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', *Utilitas* 15/3: 308–314.

Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* 69/3: 443–445.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', *Global Policy* 4/1: 15–31.

Bostrom, N. (2014), *Superintelligence: Path, Dangers, Strategies* (Oxford University Press).

Bradley, R. (2017), *Decision Theory with a Human Face* (Cambridge University Press).

Brandt, R. (1959), *Ethical Theory* (Prentice-Hall).

Bricker, D. and Ibbitson, J. (2019), *Empty Planet: The Shock of Global Population Decline* (Crown Publishing Group).

Broome, J. (2008), 'The Ethics of Climate Change', *Scientific American* 298: 96–102.

Bryson, S., Kunimoto, M., Kopparapu, R., and Zamudio, K. (2021), 'The Occurrence of Rocky Habitable-Zone Planets around Solar-Like Stars from Kepler Data', *The Astronomical Journal* 161/1: 36.

Buchak, L. (2013), *Risk and Rationality* (Oxford University Press).

Buchholz, W. and Schumacher, J. (2010), 'Discounting and Welfare Analysis Over Time: Choosing the η', *European Journal of Political Economy* 26/3: 372–385.
Bykvist, K. (2011), 'How to Do Wrong Knowingly and Get Away with It', in F. Svensson and R. Śliwiński (eds.), *Neither/Nor. Philosophical Papers Dedicated to Erik Carlson on the Occasion of His Fiftieth Birthday*. Uppsala Philosophical Studies 58. (Department of Philosophy, Uppsala University), 31–47.
Caplan, B. (2008), 'The Totalitarian Threat', in N. Bostrom and M. Cirkovic (eds.) *Global Catastrophic Risks* (Oxford University Press), 504–520.
Cassman, K. G., Dobermann, A. R., and Walters, D. T. (2002), 'Agroecosystems, Nitrogen-Use Efficiency, and Nitrogen Management', *AMBIO: A Journal of the Human Environment* 31/2: 132–140.
Chalmers, D. J. (1996), *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press).
Christiano, P. (2014), 'We Can Probably Influence the Far Future', *Rational Altruist*, https://rationalaltruist.com/2014/05/04/we-can-probably-influence-the-far-future/ (accessed 22 January 2025).
Cline, W. R. (1992), *The Economics of Global Warming* (Institute for International Economics).
Cohen, J. E. (1995), *How Many People Can the Earth Support?* (Norton).
Cowen, T. (1992), 'Consequentialism Implies a Zero Rate of Intergenerational Discount', in P. Laslett and J. S. Fishkin (eds.), *Justice Between Age Groups and Generations* (Yale University Press), 162–168.
Dasgupta, P. (2008), 'Discounting Climate Change', *Journal of Risk and Uncertainty* 37: 141–169.
Department of Transport. (2020), 'Reported Road Casualties in Great Britain: Provisional Results 2019', https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/904698/rrcgb-provisional-results-2019.pdf (accessed 28 May 2021).
Dietz, S., Hepburn, C. J., and Stern, N. (2008), 'Economics, Ethics and Climate Change', in K. Basu and R. Kanbur (eds.), *Arguments for a Better World: Essays in Honour of Amartya Sen. Volume II: Society, Institutions and Development* (Oxford University Press), 365–386.
Dorsey, D. (2012), 'Objective Morality, Subjective Morality and the Explanatory Question', *Journal of Ethics and Social Philosophy* 6/3: 1–24.
Drupp, M. A., Freeman, M., Groom B., and Nesje F. (2018), 'Discounting Disentangled', *American Economic Journal: Economic Policy* 10: 109–134.
Dubra, J., Maccheroni F., and Ok, E. A. (2004), 'Expected Utility Theory without the Completeness Axiom', *Journal of Economic Theory* 115/1: 118–133.
Elga, A. (2010), 'Subjective Probabilities Should Be Sharp', *Philosopher's Imprint* 10/5: 1–11.
Ellsberg, D. (1961), 'Risk, Ambiguity, and the Savage Axioms', *The Quarterly Journal of Economics* 75/4: 643–669.
Erev, I., Glozman, I., and Hertwig, R. (2008), 'What Impacts the Impact of Rare Events', *Journal of Risk and Uncertainty* 36/2: 153–177.
Ewing, A. C. (1948), *The Definition of Good* (Routledge & Kegan Paul).
Finnveden, L., Riedel, J., and Shulman, C. (2022), 'AGI and Lock-in', *The Effective Altruism Forum*. https://forum.effectivealtruism.org/posts/KqCybin8rtfP3qztq/agi-and-lock-in
Flyvbjerg, B. (2008), 'Curbing Optimism Bias and Strategic Misrepresentation in Planning: Reference Class Forecasting in Practice', *European Planning Studies* 16/1: 3–21.
Frick, J. (2015), 'Contractualism and Social Risk', *Philosophy & Public Affairs* 43/3: 175–223.
Galaabaatar, T. and Karni, E. (2013), 'Subjective Expected Utility with Incomplete Preferences', *Econometrica* 81/1: 255–284.
Gibbard, A. (2005), 'Truth and Correct Belief', *Philosophical Issues* 15: 338–350.
GiveWell. (2018), 'Estimating the Funding Gaps for Distribution of Antimalarial Nets and Seasonal Malaria Chemoprevention', https://www.givewell.org/international/technical/programs/malaria-funding-gaps#What_would_it_cost_to_deliver_nets_to_everyone_who_needed_them (accessed 9 February 2021).
GiveWell. (2020a), 'GiveWell's Cost-Effectiveness Analyses', https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models (accessed 26 January 2021).
GiveWell. (2020b), 'Our Criteria for Top Charities', https://www.givewell.org/how-we-work/criteria#Criteria (accessed 28 May 2021).
Gollier, C. (2013), *Pricing the Planet's Future: The Economics of Discounting in an Uncertain World* (Princeton University Press).
Good, I. J. (1966), 'Speculations Concerning the First Ultraintelligent Machine', in F. L. Alt and M. Rubinoff (eds.), *Advances in Computers* 6. (Academic Press), 31–88.
Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans O. (2018), 'When Will AI Exceed Human Performance? Evidence from AI Experts', *Journal of Artificial Intelligence Research* 62: 729–754.
Graham, P. A. (2010), 'In Defense of Objectivism about Moral Obligation', *Ethics* 121: 88–115.
Greaves, H. (2016), 'Cluelessness', *Proceedings of the Aristotelian Society* 116/3: 311–339.
Greaves, H. (2017), 'Discounting for Public Policy: A Survey', *Economics and Philosophy* 33/3: 391–439.

Greaves, H., Thomas, T., Mogensen, A., and MacAskill, W. (2024), 'On The Desire to Make a Difference', *Philosophical Studies* 181: 1599–1626.
Harrod, R. F. (1948), *Towards a Dynamic Economics* (Macmillan).
Hawking, S. (2018), *Brief Answers to the Big Questions* (John Murray Press).
Heyd, D. (1988), 'Procreation and Value: Can Ethics Deal with Futurity Problems?' *Philosophia* 18/2–3: 151–170.
Hooker, B. (2009), 'The Demandingness Objection', in T. Chappell (ed.), *The Problem of Moral Demandingness* (Palgrave McMillan), 148–162.
House of Commons Public Administration Select Committee. (2013), 'The role of the Charity Commission and "public benefit": Post-legislative scrutiny of the Charities Act 2006'. Third Report of Session 2013–14. Volume I: Report, together with formal minutes, oral and written evidence, https://publications.parliament.uk/pa/cm201314/cmselect/cmpubadm/76/7602.htm (accessed 22 January 2025).
Howard-Snyder, F. (2005), 'It's the Thought that Counts', *Utilitas* 17: 265–281.
Isaacson, W. (2003), *Benjamin Franklin: An American Life* (Simon & Schuster).
Jordà, Ò., Knoll, K., Kuvshinov, D., Schularick, M., and Taylor, A. M (2019), 'The Rate of Return on Everything, 1870–2015', *The Quarterly Journal of Economics* 134/4: 1225–1298.
Kagan, S. (1984), 'Does Consequentialism Demand Too Much? Recent Work on the Limits of Obligation', *Philosophy & Public Affairs* 13/3: 239–254.
Kahneman, D. and Lovallo, D. (1993), 'Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking', *Management Science* 39/1: 17–31.
Kahneman, D. and Tversky, A. (1979), 'Prospect Theory: An Analysis of Decision under Risk', *Econometrica* 47/2: 263–291.
Kant, I. (1996), 'On a Supposed Right to Lie from Philanthropy', in M. Gregor (transl./ed.) *Immanuel Kant: Practical Philosophy* (Cambridge University Press), 605–616.
Legg, S. (2008), *Machine Super Intelligence*, PhD thesis, University of Lugano.
Lenman, J. (2000), 'Consequentialism and Cluelessness', *Philosophy & Public Affairs* 29/4: 342–370.
Lewis, D. (1980), 'Mad Pain and Martian Pain', *Readings in the Philosophy of Psychology* 1: 216–222.
Lewis, G. (2016), 'Beware Surprising and Suspicious Convergence', *The Effective Altruism Forum* https://forum.effectivealtruism.org/posts/omoZDu8ScNbot6kXS/beware-surprising-and-suspicious-convergence (accessed 22 January 2025).
Liao, S. M. (2020), 'The Moral Status and Rights of Artificial Intelligence', in S. M. Liao (ed.), *Ethics of Artificial Intelligence* (Oxford University Press), 480–498.
MacAskill, W. (2014), 'Replaceability, Career Choice, and Making a Difference', *Ethical Theory and Moral Practice* 17: 269–283.
MacAskill, W. (2019), 'When Should an Effective Altruist Donate?', GPI Working Paper 8-2019 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/william-macaskill-when-should-an-effective-altruist-donate/ (accessed 22 January 2025).
MacAskill, W. (2022), *What We Owe The Future* (Basic Books).
MacAskill, W., Mogensen, A., and Ord, T. (2018), 'Giving Isn't Demanding', in P. Woodruff (ed.), *The Ethics of Giving: Philosophers' Perspectives on Philanthropy* (Oxford University Press), 178–203.
Machina, M. J. and Siniscalchi, M. (2014), 'Ambiguity and Ambiguity Aversion', in *Handbook of the Economics of Risk and Uncertainty, Volume 1* (North Holland), 729–807.
Mason, E. (2013), 'Objectivism and Prospectivism about Rightness', *Journal of Ethics and Social Philosophy* 7/2: 1–21.
Meacham, C. (2012), 'Person-Affecting Views and Saturating Counterpart Relations', *Philosophical Studies* 158/2: 257–287.
Meng, F. and Ellis, T. (2020), 'The Second Decade of Synthetic Biology: 2010–2020', *Nature Communications* 11: 5174.
Millett, P. and Snyder-Beattie, A. (2017), 'Existential Risk and Cost-Effective Biosecurity', *Health Security* 15/4: 373–383.
Mogensen, A. (2019), 'The Only Ethical Argument for Positive ?' GPI Working Paper 5-2019 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/andreas-mogensen-the-only-ethical-argument-for-positive-delta-2/ (accessed 22 January 2025).
Mogensen, A. (2021), 'Moral Demands and the Far Future', *Philosophy and Phenomenological Research* 101/3: 567–585.
Moore, G. E. (1903), *Principia Ethica* (Cambridge University Press).
Mulgan, T. (2001), *The Demands of Consequentialism* (Oxford University Press).

Musk, E. (2014), 'One-on-one with Elon Musk. Talk at the MIT AeroAstro 1914-2014 Centennial Symposium', https://aeroastro.mit.edu/videos/centennial-symposium-one-one-one-elon-musk (accessed 22 January 2025).
Nagel, T. (1978), 'Ruthlessness and Public Life', in S. Hampshire (ed.), *Public and Private Morality* (Cambridge University Press), 75–91.
Newberry, T. (2021a), 'How Many Lives Does the Future hold?', GPI Technical Report T2-2021 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/how-many-lives-does-the-future-hold-toby-newberry-future-of-humanity-institute-university-of-oxford/ (accessed 22 January 2025).
Newberry, T. (2021b), 'How Cost-Effective Are Efforts to Detect Near-Earth-Objects?', GPI Technical Report T1-2021 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/how-cost-effective-are-efforts-to-detect-near-earth-objects-toby-newberry-future-of-humanity-institute-university-of-oxford/ (accessed 22 January 2025).
Ngo, R. (2020), 'AGI Safety from First Principles', https://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view (accessed 22 January 2025).
O'Donoghue, T. and Somerville, J. (2018), 'Modeling Risk Aversion in Economics', *The Journal of Economic Perspectives* 32/2: 91–114.
Ochs, A. (2019), 'What Is the Oldest Charitable Trust in the U.S., and What Does It Fund?', *Inside Philanthropy*, https://www.insidephilanthropy.com/home/2019/4/21/what-is-the-oldest-charitable-trust-in-the-us-and-what-does-it-fund (accessed 22 January 2025).
Olsen, K. (2017), 'A Defense of the Objective/Subjective Moral Ought Distinction', *The Journal of Ethics* 21/4: 351–373.
Open Philanthropy Project. (2020), Grants database, https://www.openphilanthropy.org/giving/grants (accessed 26 January 2021).
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Parfit, D. (1984), *Reasons and Persons* (Clarendon Press).
Parfit, D. (2011), *On What Matters: Volume 1*, ed. S. Scheffler (Oxford University Press).
Pierson, P. (2000), 'Increasing Returns, Path Dependence, and the Study of Politics', *The American Political Science Review* 94/2: 251–267.
Pigou, A. C. (1932), *The Economics of Welfare: Volume 1*, 4th edition (Macmillan).
Portmore, D. W. (2011), *Commonsense Consequentialism: Wherein Morality Meets Rationality* (Oxford University Press).
Posner, R. (2004), *Catastrophe: Risk and Response* (Oxford University Press).
Pratt, J. W. (1964), 'Risk Aversion in the Small and in the Large', *Econometrica* 32/1–2: 122–136.
Prichard, H. A. (1932) [2002]), 'Duty and Ignorance of Fact', in J. McAdam (ed.), *Moral Writings* (Oxford University Press), 85–110.
Quiggin, J. (1982), 'A Theory of Anticipated Utility', *Journal of Economic Behavior and Organization* 3: 323–343.
Quiggin, J. and Wakker, P. (1994), 'The Axiomatic Basis of Anticipated Utility: A Clarification', *Journal of Economic Theory* 64: 486–499.
Rabin, M. (2000), 'Risk Aversion and Expected-Utility Theory: A Calibration Theorem', *Econometrica* 68/5: 1281–1292.
Ramsey, F. P. (1928), 'A Mathematical Theory of Saving', *Economic Journal* 38: 543–559. (Reprinted in F. P. Ramsey, *Foundations: Essays in Philosophy, Logic, Mathematics, and Economics*, ed. D. H. Mellor.)
Rees, M. (2018), *On the Future: Prospects for Humanity* (Princeton University Press).
Reese, J. (2018), 'Why I Prioritize Moral Circle Expansion over Artificial Intelligence Alignment', *Effective Altruism Forum*, https://forum.effectivealtruism.org/posts/BY8gXSpGijypbGitT/why-i-prioritize-moral-circle-expansion-over-artificial (accessed 9 February 2021).
Ross, W. D. (1930), *The Right and the Good* (Oxford University Press).
Ross, W. D. (1939), *The Foundations of Ethics* (Clarendon Press).
Rowe, T. and Voorhoeve, A. (2018), 'Egalitarianism under Severe Uncertainty', *Philosophy & Public Affairs* 46/3: 239–268.
Russell, B. (1966), 'The Elements of Ethics', in *Philosophical Essays* (Simon and Schuster), 13–59.
Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking).
Sackmann, I.-J., Boothroyd, A. I., and Kraemer, K. E. (1993), 'Our Sun. III. Present and Future', *The Astrophysics Journal* 418: 457–468.
Sandberg, A. (2013), 'Feasibility of Whole Brain Emulation', in V. C. Müller (ed.), *Theory and Philosophy of Artificial Intelligence* (Springer), 251–64.

Sandberg, A. (2014), 'Ethics of Brain Emulations', *Journal of Experimental and Theoretical Artificial Intelligence* 26/3: 439–457.
Sandberg, A. and Bostrom, N. (2008a), 'Global Catastrophic Risks Survey', Future of Humanity Institute Technical Report #2008-1, https://www.fhi.ox.ac.uk/reports/2008-1.pdf (accessed 22 January 2025).
Sandberg, A. and Bostrom N. (2008b), 'Whole Brain Emulation: A Roadmap', Future of Humanity Institute Technical Report #2008-3, https://www.fhi.ox.ac.uk/reports/2008-3.pdf (accessed 22 January 2025).
Sarbu, I. and Sebarchievici, C. (2017), *Solar Heating and Cooling Systems: Fundamentals, Experiments and Applications* (Elsevier).
Scanlon, T. M. (1998), *What We Owe to Each Other* (Harvard University Press).
Schröder, K.-P. and Smith, R. C. (2008), 'Distant Future of the Sun and Earth Revisited', *Monthly Notices of the Royal Astronomical Society* 386/1: 155–163.
Setiya, K. (2014), 'The Ethics of Existence', *Philosophical Perspectives* 28/1: 291–301.
Shulman, C. (2020), 'What Do Historical Statistics Teach Us about the Accidental Release of Pandemic Bioweapons?', *Reflective Disequilibrium*, http://reflectivedisequilibrium.blogspot.com/2020/10/what-do-historical-statistics-teach-us.html (accessed 22 January 2025).
Sidgwick, H. (1890), *The Methods of Ethics* (Macmillan).
Simon, J. (1998), *The Ultimate Resource 2*, revised edition (Princeton University Press).
Singer, P. (1972), 'Famine, Affluence and Morality', *Philosophy & Public Affairs* 1/3: 229–243.
Snyder-Beattie, A., Ord, T., and Bonsall M. B. (2019), 'An Upper Bound for the Background Rate of Human Extinction', *Scientific Reports* 9: 11054.
Solow, R. M. (1974), 'The Economics of Resources or the Resources of Economics', *American Economic Review Papers and Proceedings* 64: 1–14.
Statt, N. (2015), 'Bill Gates Is Worried about Artificial Intelligence Too', *CNET*, https://www.cnet.com/news/bill-gates-is-worried-about-artificial-intelligence-too/ (accessed 22 January 2025).
Steele, K. and Stefánsson, H. O. (2021), *Beyond Uncertainty: Reasoning with Unknown Possibilities* (Cambridge University Press).
Stefánsson, H. O. and Bradley, R. (2019), 'What Is Risk Aversion?', *The British Journal for the Philosophy of Science* 70/1: 77–102.
Stern, N. (2007), *The Economics of Climate Change* (Cambridge University Press).
Stix, M. (2002), *The Sun: An Introduction* (Springer).
Tarsney, C. (2019), 'The Epistemic Challenge to Longtermism', GPI Working Paper 10-2019 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/christian-tarsney-the-epistemic-challenge-to-longtermism/
Tegmark, M. (2017), *Life 3.0: Being Human in the Age of Artificial Intelligence* (Knopf).
Texas Education Agency. (2020), Texas Permanent School Fund https://tea.texas.gov/finance-and-grants/texas-permanent-school-fund (accessed 22 January 2025).
Thomas, T. (2019), 'The Asymmetry and the Long Term', GPI Working Paper 11-2019 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/teruji-thomas-the-asymmetry-uncertainty-and-the-long-term/ (accessed 22 January 2025).
Thomson, J. J. (1986), 'Imposing Risks', in W. Parent (ed.), *Rights, Restitution, and Risks* (Harvard University Press), 173–191.
Thomson, J. J. (2008), *Normativity* (Open Court).
Trammell, P. (2020), 'Dynamic Public Good Provision under Time Preference Heterogeneity: Theory and Applications to Philanthropy', https://philiptrammell.com/static/PatienceAndPhilanthropy.pdf (accessed 22 January 2025).
Trautmann, S. T. and van de Kuilen, G. (2015), 'Ambiguity Attitudes', in G. Keren and G. Wu (eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making, I.* (Wiley), 89–116.
Turing, A. (1951), 'Can Digital Computers Think?' *BBC Radio*, May 15 1951. https://www.bbc.co.uk/videos/cqvv8n1d8dxo
United Nations, Department of Economic and Social Affairs, Population Division. (2019), *World Population Prospects 2019: Highlights*, World Population Prospects 2019 Highlights (accessed 22 January 2025).
University of Texas System. (2021), 'The Permanent University Fund (PUF)', https://www.utsystem.edu/puf (accessed 22 January 2025).
Van Den Bergh, J. C. J. M. and Rietveld, P. (2004), 'Reconsidering the Limits to World Population: Meta-analysis and Meta-prediction', *BioScience* 54/3: 195–204.
Voorhoeve, A. (2014), 'How Should We Aggregate Competing Claims?' *Ethics* 125: 64–87.
Wolf, E. and Toon, O. B. (2015), 'The Evolution of Habitable Climates under the Brightening Sun', *Journal of Geophysical Research: Atmospheres* 120: 5775–5794.

Yudkowsky, E. (2013), 'Intelligence Explosion Microeconomics', Technical Report 2013-1 (Machine Intelligence Research Institute).
Zimmerman, M. J. (2006), 'Is Moral Obligation Objective or Subjective?', *Utilitas* 18: 329–61.
Zimmerman, M. J. (2008), *Living with Uncertainty: The Moral Significance of Ignorance* (Cambridge University Press).

3

Longtermism and Neutrality about More Lives

Katie Steele

1 Introduction

The momentum of the so-called *Longtermism* movement—escalated by the publication of William MacAskill's (2022) book *What We Owe the Future*—owes in part to an exciting vision of the power of the present generation to give humanity a leg-up towards a brighter, longer-lasting future. The headline calls to action are for significant investments to reduce the risk of what many perceive as 'futuristic threats', for example, premature human extinction due to engineered pathogens, or a long-enduring tyranny enabled by artificial superintelligence (ASI). The idea is that these are the sorts of threats to human flourishing that, while not necessarily immediate, span the longest timescales and are most impactful. Indeed, the timescales are so long that we supposedly have strong moral reason to pursue even small reductions in the risk of these threats materialising.

While a gripping vision, many find the idea of devoting significant resources to addressing future threats deeply troubling if it means diverting resources from, say, tackling the more immediate and cumulative problem of climate change, or reforming existing institutions to reduce economic inequalities and other injustices. I too find this troubling. Some simply dismiss concerns about future threats on the grounds that they rely on a faulty premise regarding the fundamental nature of moral reasons directed at human welfare. The supposed error lies in treating more worthwhile lives as inherently better since this increases *total* positive human welfare. The proposed alternative is that more worthwhile lives is neither better nor worse, but rather inherently *neutral*. Kieran Setiya[1], for example, has made this claim as follows:

> But if neutrality is right, the longtermist's mathematics rest on a mistake: the extra lives don't make the world a better place, all by themselves. Our ethical equations are not swamped by small risks of extinction. And while we may be doing much less than we should to address the risk of a lethal pandemic, value lock-in, or nuclear war, the truth is much closer to common sense than MacAskill would have us believe. We should care about making the lives of those who will exist better, or about the fate of those who will be worse off, not about the number of good lives there will be.

One problem with this statement is that it groups different kinds of future threats together—both those that affect the number of worthwhile lives (premature human extinction) and those that do not affect the number of lives at all, just the quality of those lives

[1] This is from his (2022) article in *Boston Review*.

Katie Steele, *Longtermism and Neutrality about More Lives* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Katie Steele 2025. DOI: 10.1093/9780191979972.003.0003

('value lock-in', or a long-enduring tyranny). But consider just the former kind of threat, for which the question of neutrality is pertinent. If the diagnosis stated above were right, we would not need to engage with empirical claims about the probability of premature human extinction, given various courses of action. Nor would we need to consider how to negotiate this uncertainty in decision making.

But unfortunately this short-circuit strategy does not work. I will argue that conclusions that we have moral reason to reduce the risk of future threats—call these *longtermist* conclusions—do not rest precariously on a deeply divisive view about the fundamental nature of moral reasons directed at human welfare. That holds for all kinds of future threats, including those that turn on premature human extinction. It is not just that there are welfarist reasons, whether totalist or neutral, to care about reductions in the risk of *specific* kinds of future threats—in particular those which affect the welfare of a population of fixed size.[2] Rather, there are welfarist reasons, whether totalist or neutral, to care about reductions in the risk of *all* kinds of future threats—including those which determine the size of the population. Moreover, I will argue that the neutral approach may lead to problematic *non-longtermist* conclusions in cases where additional lives are worthwhile and yet fall short of a utopian level of welfare.

This does not mean that longtermist conclusions—that we should not merely *care* about future threats but have moral reason to act to reduce their risk—are sound. To effectively challenge them, however, we must instead redirect attention to the *non-moral* premises on which their robust appeal—given more fine-grained disagreements about welfare—depends. It is one thing to care about the *possibility* of a reduction in the risk of a future threat. It is quite another to take there to be strong moral reason to *actually* pursue, in a direct way, a reduction in the risk of that threat. This point may be clear enough in the case of future threats to a fixed population. I therefore start out—in the next section—with this case. We see that the empirical premises are crucial to the choice conclusions. With select premises, differing accounts of moral reasons directed at human welfare, whether totalist or neutral, converge on seemingly radical choice conclusions. The remainder of the chapter argues that the case of reducing the risk of premature extinction is not much different; seemingly radical choice conclusions to this end are not greatly sensitive to the difference between totalist and neutral approaches to moral reasons directed at human welfare. That is, here too, totalism in the 'longtermist's mathematics' plays a relatively minor role.

2 Future threats to a fixed population

First some set-up that will assist the discussion throughout the chapter. We are concerned with the deliberations of a decision maker who faces choices with far-reaching implications. This may be a single individual or a governing body. And we will consider just the

[2] That is the line pursued by Hilary Greaves and William MacAskill (this volume) in their brief discussion of the sensitivity of longtermist conclusions to the details of moral theory. They say, cautiously: 'According to the spirit of a person-affecting approach, premature extinction is in itself at worst neutral: if humanity goes prematurely extinct, then there does not exist any person who is worse off as a result of that extinction, and, according to a person-affecting principle, it follows that the resulting state of affairs is not worse. The far-future benefits of extinction risk mitigation may therefore beat the best near-future benefits only conditional on controversial population axiologies.'

Table 3.1 Choice problem for a fixed-sized population.

	State 1		State 2	
	near	*far*	**near**	**far**
short-sighted	*modest*	*none*	modest	none
far-sighted	*none*	*extreme*	none	none

moral reasons directed at general human welfare—what might be called reasons of *impartial beneficence*, but which we will refer to simply as 'welfarist reasons'—that bear on this decision maker's choice. There may be other moral reasons, and perhaps more important ones, that bear on her choice. Perhaps it is reasonable to assume that all else is equal with respect to these other moral reasons, which might concern special relationships or rights and obligations. But perhaps that is not a reasonable assumption for the kinds of choices under discussion. Our decision maker may in any case have further non-moral reasons for choice, for instance, personal or prudential reasons. There is an important and overarching question of how *all* her reasons, both moral and non-moral, may be weighed against each other. But we put this overarching question aside here. Our discussion is limited to the welfarist reasons that bear on a decision maker's choice, as this is arguably how longtermist conclusions are best interpreted.

Consider the claim that our decision maker has strong welfarist reasons to pursue a 'far-sighted' option pertaining to the use of some resource X that would exclusively benefit the many who will exist in the further future. For instance, the option in question might involve spending X on reducing the risk of debilitating 'value lock-in' or long-lasting tyranny enabled, say, by artificial superintelligence (ASI). That our decision maker has strong welfarist reasons to pursue this option depends on a number of premises, regarding: (i) the empirical decision set-up, (ii) the proper resolution of uncertainty, and (iii) the nature of the welfarist reasons. Note that one can think of the welfarist reasons as tracking the relative moral goodness of the outcomes with respect to human welfare (or the 'welfarist good' for short).[3]

Table 3.1 exemplifies the substantive nature of the first kind of premises. The table depicts just two competing options for spending X, or two deviations from the status quo option (which, while not explicitly represented, serves as the reference point for describing the outcomes of the other options). All of the possible outcomes contain populations of the same size.[4] The outcome of the 'far-sighted' option, unlike the 'short-sighted' option, depends on which of the possible states of the world turns out to be actual. The first state in the table, 'State 1', is highlighted to emphasise that the welfarist comparison of the options depends primarily on the outcomes under this state. This is the state of the world in which

[3] One can interpret 'welfarist good' in a very general way; the associated ranking of outcomes may, for instance, be 'context sensitive' such that *transitivity* is not satisfied across choice sets. That said, I will later assume transitivity in my discussion of the implications of neutrality in section 4.

[4] We might add an even stronger premise that all of the possible outcomes contain populations constituted by the same people. Then the choice recommended by welfarist reasons would be even less sensitive to differences in the various accounts of the welfarist good. But since the focus here is the fundamental difference between totalist and neutral approaches to the value of additional worthwhile lives, we will put aside more extreme 'person-affecting' versions of the latter, whereby the relative goodness of a pair of outcomes depends only on their welfare effects with respect to those who exist on both outcomes.

pursuing the 'far-sighted' option makes a significant difference to the welfare of the 'far' group, which contains a relatively enormous number of people further away in time—in aggregate, this is an extremely positive change in welfare from the status quo. Even if we assume that State 1 is extremely unlikely compared to State 2, the extraordinary welfarist difference in the outcomes under State 1, let us say, is such that the relatively minor welfarist difference in the outcomes under State 2 is not important.

This brings us to the question of what our decision maker has most welfarist reason to do. If it is assumed that she faces the choice problem represented by Table 3.1, then it is not such a large jump to the conclusion that she has (even strong) welfarist reason to pursue the 'far-sighted' option, given standard ways of resolving uncertainty.[5] This conclusion does not depend at all on the difference between the totalist and neutral approaches to the value of additional worthwhile human lives, given that the size of the population is not affected by this choice. Just *how robust* is the conclusion that our decision maker has most welfarist reason to pursue the far-sighted option depends on some further considerations that are elided in Table 3.1. For instance, it will be more robust if the many further away in time are the worst off and the welfare increments for each of these individuals in State 1 are moreover large enough to count as 'relevant claims'[6] when compared to the welfare increments at stake for each of the individuals nearby in time. So there is some slightly more detailed specification of Table 3.1 such that, according to a large range of views about the relative welfarist goodness of outcomes, our decision maker has (even strong) welfarist reason to pursue the far-sighted option.

The argument that we have strong welfarist reason to pursue a reduction in the risk of a long-enduring tyrannical regime enabled by ASI is puzzling and yet compelling when presented along the lines of Table 3.1. It is puzzling because it is counter to ordinary practice to think that our welfarist reasons direct us to focus on the far future. And yet it is compelling because the conclusion seems irresistible since robust to differences in views about the welfarist good. But the conclusion is only irresistible on the assumption that the decision problem really looks something like Table 3.1. This seems initially plausible, if we consider the possibility, however remote, of a long-enduring tyrannical regime. The aggregate welfare benefit of reducing the risk of this outcome, even very slightly, is presumably large enough to easily outweigh any modest welfare benefits that would accrue to relatively few people nearby in time. But a decision problem like that in Table 3.1 depends on further empirical premises: that acting to directly reduce the risk of specific future threats will really work, and that the opportunity costs, or the relative benefits of alternative actions, are located just within the nearby time period and are thus relatively small.

In the next section, I claim that the same sorts of considerations apply to arguments that we have strong welfarist reasons to pursue a reduction in the risk of premature human extinction. Here it may seem that the empirical details matter less; that one can get off the

[5] While there is not the space to properly address the resolution of uncertainty in this setting, Table 3.1 facilitates an *ex post* approach, which involves evaluating the respective outcomes and then using this information, as well as the probability of the outcomes, in evaluating the options. A 'standard way' of evaluating the options would be one that conforms to (some conservative generalisation of) the expected value principle, whereby the higher the expected value, the better the option.

[6] For an influential welfarist view that turns on aggregating only 'relevant claims', see Alex Voorhoeve's (2014) proposal.

longtermist wagon simply by denying the totalist approach to the value of additional worthwhile human lives. But we will see that this is a mistake.

3 Future threats to population size

Let's for the moment assume a totalist approach to the value of additional human lives. On such an approach, every extra life with positive welfare—assuming welfare is scaled so that a life of zero welfare is neither good nor bad for the person living it—increases the welfarist good of an outcome, and every extra life with negative welfare decreases the welfarist good of an outcome. (For the specific totalist approach known as *total utilitarianism*, for instance, the overall welfarist good of an outcome is just the sum of the welfare of all who exist in that outcome.) It is not hard to see that, on a totalist account of the welfarist good, greatly increasing the number of persons with positive welfare who will exist is, all else equal, a very good thing to do.

Accordingly, it seems *prima facie* plausible that we have strong welfarist reasons, if some version of totalism is right, to reduce the risk of premature human extinction. After all, premature human extinction cuts off an indefinitely long period of human history that plausibly contains on balance extraordinary amounts of positive welfare. Just like in the case of a long-enduring tyranny, the welfare benefit of reducing the risk of premature human extinction, even very slightly, seems then large enough to easily outweigh the welfare benefit of any shorter-term project.

That is, it seems that our plight as decision makers may plausibly resemble Table 3.2, where the 'far-sighted option' is to spend some resource X on directly trying to reduce the risk of premature human extinction. And if so, we would have strong welfarist reasons, if these reasons are broadly totalist at least, to pursue this option. As before, the table depicts the outcomes for two options or deviations from the status quo. The outcomes are presented with respect to a partition of the possible overall population: in this case, the first group contains some inevitable number of people, the 'base-sized' group, and the second group contains some further number of people that is contingent on the natural and agential forces pertinent to this choice problem, that is, the 'spillover-sized' group.[7] Again, 'State 1' is highlighted. This is the state of the world in which pursuing the 'far-sighted' option makes a significant difference to the welfare of the 'spillover-sized' group, because it adds an enormous number of people to this group.

Many are unmoved, however, by the use of Table 3.2, or something like it, to argue that there are strong welfarist reasons to directly aim at reducing the risk of premature human

Table 3.2 Choice problem for a varying-sized population.

	State 1		State 2	
	base	***spillover***	**base**	**spillover**
short-sighted	*modest*	*none*	modest	none
far-sighted	*none*	*extreme*	none	none

[7] This is a variant of the 'necessary people' versus 'possible people' distinction. There is a subtle difference in that here we are focusing just on the size of the population. By contrast, the 'necessary people', for instance, are the specific people who will exist, whatever choice one makes; they are not merely those who make up some fixed number of people who will exist, whatever choice one makes.

extinction. The thought is that any such argument is a non-starter since it depends precariously on welfarist reasons of a totalist kind. The alternative approach to welfarist reasons that many find attractive, especially in view of future threats that turn on premature human extinction, is what was referred to above as the neutral approach, or 'neutrality' for short. On this approach, extra lives with positive welfare are neither good nor bad in and of themselves, but rather are neutral.[8] It does seem that welfarist reasons of a neutral kind would provide no reason to aim at ensuring the existence of more people, especially when there are significant opportunity costs to presently existing people, or more generally to a population of guaranteed size. Only under totalist approaches is there an extraordinary difference in welfare between the two outcomes under 'State 1', a difference that effectively determines the welfarist comparison of the options even if 'State 1' has extremely small probability.

Before going on in the next section to examine the implications of neutrality, I will pause to state the position more carefully. The thought is that there is some range of welfare that we can refer to as the *neutral range*, whereby adding an extra person whose welfare falls within that range does not make the world better or worse in and of itself. The neutral range is typically thought to have a lower bound equating to the value of a life that is neither good nor bad (zero, by convention). Lives with welfare below the neutral range—those that are bad for the person living them—are thought to detract from the welfarist goodness of an outcome.[9] The upper bound of the neutral range could be very high or even infinite. Arguably, the common thought is that any good life, no matter how good, does not in itself make the world a better place. Wlodek Rabinowicz (2022, 116) refers to this as the 'radical' interpretation of neutrality, whereby the neutral range of welfare extends from zero all the way up to the value of a maximally good life, or else infinity if there is no upper bound on a person's welfare. (He contrasts this with the 'moderate' interpretation, whereby the neutral range of welfare extends from zero to some positive, not-too-high level of welfare.) Here we will initially stick with the 'radical' interpretation, since, in spite of Rabinowicz's label, that is arguably what many advocates of neutrality have in mind.

Neutrality seems initially to have much to recommend it. Indeed, Jacob Nebel (2019) and Johann Frick (2020) have provided rich accounts of the notion of value that underpins intuitions supporting neutrality. The idea is that positive welfare is not something we have *unconditional* reason to bring about. Nebel and Frick propose instead welfare-related reasons that are *conditional* on the person in question's existence, or the *bearer* of welfare. This leads to an account whereby outcomes are ranked according to a notion of conditional welfare value: It is better that bearers of welfare have positive rather than negative welfare (and more generally higher rather than lower welfare), and it is better for there to be no bearer of welfare than for there to be one with negative welfare, but it is neither better nor worse for there to be no bearer of welfare than for there to be one with positive welfare. Surely then, one might think, preventing premature human extinction (or rather, increasing the number of lives with positive welfare) is not worth any sacrifice if one assumes neutrality.

[8] Not everyone who endorses neutrality thinks it is obviously a mistake to worry about premature human extinction. For instance, Frick's (2017) project is to offer an account, consistent with neutrality, of our reasons to prevent premature human extinction.

[9] That is, neutrality is typically spelled out in a way that is sensitive to the so-called 'Procreation Asymmetry', whereby there is a duty not to bring into existence a person with a bad life, presumably because this would be bad, and yet there is no duty to bring into existence a person with a good life, presumably because this is neither good nor bad.

4 Implications of neutrality

The implications of neutrality turn out to be difficult to glean, however, just by contemplating the *kind* of welfare value that underpins this approach. John Broome (2004; 2005) showed that the resulting ranking of outcomes with respect to welfare (or 'welfarist ranking') is more complicated than first appearances suggest. Here I will extend Broome's insights to outcomes with long time horizons that allow for much variation in the number of persons who exist. We will see that neutrality does not in fact preclude acting on welfarist reasons to directly reduce the risk of premature human extinction, even at great opportunity cost to those living in the present, or more generally to the fixed-size population that is bound to obtain. In fact, neutrality does not preclude even more counter-intuitive conclusions.

4.1 Greedy neutrality

The starting point for Broome (2004; 2005) is what precisely is the relationship, in terms of the welfarist good, between some original population and that population with an additional life of positive welfare? Neutrality says that the former, with the additional life, is neither better nor worse than the latter. Should this be interpreted as indifference? Arguably not, since if the indifference and better-than relations satisfy *transitivity*, we get an inconsistency with the *(strong) Pareto principle*. Consider the following case, where the population outcomes, *A*, *B*, and *C*, are represented as vectors. Each entry in the vector corresponds to a particular person and gives their lifetime welfare, with '–' representing that the person does not exist on that outcome:

$$A = (2,3,-)$$
$$B = (2,3,4)$$
$$C = (2,3,5)$$

On the indifference interpretation, neutrality here implies indifference between *A* and *B* ($A = B$) and between *A* and *C* ($A = C$). But then by transitivity of indifference, we would get $B = C$. But *C* weakly Pareto dominates *B*, so by the (strong) Pareto principle *C* is better than *B* ($C \succ B$). Hence, an inconsistency.

Broome (2004; 2005) proposes instead that neutrality be fleshed out in terms of the relation of incommensurability. In that way, neutrality conforms with both strong Pareto and transitivity of the better-than relation with respect to the welfarist good. For our example, *A* is then incommensurable with both *B* and *C* ($A \approx B$ and $A \approx C$). Since incommensurability is not transitive, we may yet have $C \succ B$. Indeed, the classic examples used to illustrate the difference between incommensurability and indifference have precisely this form. For instance, in the ranking of holiday destinations, to say that a rugged wilderness escape, *A*′, and a cultural city sojourn, *B*′, are incommensurable (that is, $A' \approx B'$) is to say that there is some enhancement (say a fixed amount of extra spending money) to the cultural city sojourn (yielding *C*′, where $C' \succ B'$), such that *A*′ and *C*′ are also incommensurable (that is, $A' \approx C'$).

What Broome further draws attention to—our starting point for the next section—is that the incommensurability of neutrality is 'greedy'. It allows good and bad changes to the original population to be offset by the addition of lives with positive welfare to the original population. (What is meant by 'offset' here is that, on net, there are no welfare reasons for or against the new augmented population involving gains or losses to the original population, as compared to the original population, since the two are incommensurable.) A simple example (owing to Rabinowicz 2009) will suffice to demonstrate both aspects of this point:

$$D = (3,4,-)$$
$$E = (3,4,1)$$
$$F = (3,3,3)$$
$$G = (3,3,-)$$

According to neutrality, D and E are incommensurable with respect to the welfarist good ($D \approx E$) and similarly F and G are incommensurable ($F \approx G$). But plausibly $F \succ E$, since F has greater equality, and also greater total welfare. But since E is not worse than D, F cannot be worse than D. And yet the move from D to F involves a decrease in welfare for someone and the addition of a new person with positive welfare. So we see that neutrality entails that the addition of a person with positive welfare (in the neutral range) can offset a decrease in welfare for those in the original population. (Broome says rhetorically that neutrality about added lives can 'swallow up' the original people's loss in welfare, neutralising it, and making the change overall not bad.) Now consider G and E. Since, according to neutrality, F is not better than G, and F is itself better than E, then E cannot be better than G. But the move from G to E involves an increase in welfare for someone and the addition of a new person with positive welfare. So we see that neutrality entails that the addition of a person with positive welfare (in the neutral range) can offset an increase in welfare for those in the original population. (Again, we might say that neutrality about added lives can 'swallow up' the original people's gains, neutralising them, and making the change overall not good.)

Broome (2004; 2005) claims that the appeal of neutrality depends on there being no swallowing up of this sort, and so the incommensurability interpretation of neutrality shows this approach to be based on some false hope. Others disagree (e.g., Rabinowicz 2009; Frick 2017); they suggest that greediness is not necessarily an unwelcome consequence of neutrality for those who find it appealing. In any case, the focus here is to simply extend Broome's investigations of the greediness of neutrality to the long-term or large-scale setting where many lives may be at stake. We will ultimately compare, in this setting, the neutral and totalist approaches to the welfarist good.

Note that our discussion will treat transitivity as a constraint on the better-than relation with respect to the welfarist good. (Further assumptions will also be introduced for ease of explication, more on which in section 4.2.) On a broader notion of the welfarist good—one that simply tracks reasons for choice in any given choice context—the associated 'more-choice-worthy-than' relation need not be transitive across choice contexts. For instance, Frick (2022) allows that the choice-worthiness relation between pairs of options may change depending on the context: for the above options, E may be no less choice-worthy than D in a pairwise comparison, but less choice-worthy where the option set also includes F, because in that case the choice of E would involve unnecessary inequality (compared to F). This sort of context dependence *may* make neutrality less greedy overall, but

it depends on the details of the fully formulated account. Teruji Thomas (2023, secs. 5.1 and 5.2) offers fully formulated accounts along these lines that he claims are defensible in the context of uncertainty, but, on these accounts, as per my discussion below, the lives of additional people matter in the welfarist comparison of options. So I tentatively suggest that weakening the transitivity assumption would not dramatically change the overall story in what follows in the ways that defenders of neutrality who wish to resist longtermist conclusions might hope.

4.2 Greedy neutrality and longtermism

Let us turn then to the long term or large scale. Even if the greediness of neutrality is not particularly surprising or unwelcome on a small scale, things may look different on a large scale. The investigation of the latter setting is, however, not so straightforward. The simple demonstrations of the greediness of neutrality, like that above, depend on very few assumptions regarding the welfarist ranking of options. When it comes to settings in which many lives are at stake—including decision problems for which the choice of option affects the threat of premature human extinction—we must introduce further assumptions to assess the implications of neutrality. Indeed, it helps to work with a rather specific account of the welfarist good and leave it largely to the reader to consider how the results would change for nearby accounts. At the very least, we can say that the results demonstrated here are not *precluded* by neutrality.

For the sake of clear explication, assume then that the welfarist good has the following features: (i) populations of differing size are compared in a way that conforms to neutrality, whereby adding lives with positive welfare to an original population yields an augmented population that is incommensurable with that original population, and (ii) populations of the same size are compared in terms of average welfare, the higher the better.[10] (The second feature is clearly arbitrary in the context of our discussion; it might just as well be some other approach to comparing populations of the same size.)

This notion of the welfarist good makes clear that neutrality does not commit one to a view whereby populations of the same size constituted by different people are incommensurable. On the account just outlined, such populations are ranked according to their average welfare. So the following statement from MacAskill (2022, 175) is rather misleading:

> Consider two people, Alice and Bob. If we keep fossil fuel subsidies, Alice will be born in 2070. If we end fossil fuel subsidies, Alice will not be born and Bob will be born instead. Both have happy lives, but, because climate change will be less extreme without fossil fuel subsidies, Bob will be happier than Alice would have been. According to the intuition of neutrality, we do not have reason to ensure that Bob exists rather than Alice. According to the intuition of neutrality, preventing Alice's existence is neither good nor bad, and

[10] This may well be an alternative description, or at least a partial description, of a specific version of 'critical-range' utilitarianism (see Rabinowicz, e.g., 2009; 2022), in which the 'critical range' extends from zero (representing the welfare of a life that is neither good nor bad for the person living it) to positive infinity.

bringing Bob into existence is also neither good nor bad. So doing both at once is neither good nor bad.

MacAskill describes a violation of what Derek Parfit (1984, 367) calls the 'No Difference View', since the change in identity here between the happy and happier person *does* affect the comparison of the two outcomes. But neutrality does not itself imply violations of No Difference. In fact, defenders of neutrality typically also defend No Difference (see, e.g., Frick 2020). For the example above, that would mean that ending fossil fuel subsidies is the better option, since were it not for a change in identity from Alice to Bob, this option weakly Pareto dominates the other. The change in identity does not affect this ranking.

So, neutrality does not in itself lead to so much incommensurability in the comparison of options as MacAskill suggests. But it still leads to a lot of troubling incommensurability owing to the greediness of neutrality.

Consider the case where added lives 'swallow up' suffering or bad changes for the original population. The recipe for generating such cases of swallowing up is as follows: Take an 'original population'. Add n lives right at the cusp of being worth living, or just above zero welfare. Call this the 'augmented population'. It is incommensurable with the original population with respect to the welfarist good. Now there will be various populations of the same size as the augmented population that are better than that population. These '+n-populations' are thus not worse than the original population; they are either incommensurable with or better than the original population. But some of these +n-populations will involve welfare losses for the original population. These are the populations in which the added lives 'swallow up' losses to the original population. For our account of the welfarist good in which same-sized populations are compared in terms of their average welfare, the +n-populations will include populations with any given sum of welfare loss to the original population; this loss must be effectively compensated by at least as great a sum welfare gain for the added n people, compared to what they each could have had in the augmented population, which was close to zero welfare.

The swallowing up gets even worse. It is not just that any welfare loss to the original population can be swallowed up by sufficient total gains for the added people relative to their lives being only just worth living (or close to zero welfare). Worse, the more people added to the original population, or the greater that n is, the less their respective welfare levels need to surpass zero for their sum gains to be sufficient.

Return then to the choice problem in Table 3.2, where under State 1, the 'far-sighted' option results in many more lives with positive welfare. On a totalist approach to the welfarist good, the 'far-sighted' option is better than the 'short-sighted' option, provided the difference in the goodness of the outcomes under State 1 is sufficiently large relative to the (very small) probability that State 1 is true. Can we get around this uncomfortable conclusion with a neutral approach to the welfarist good? Neutrality does not preclude the 'far-sighted' option being incommensurable with the 'short-sighted' option. This is to say there is no welfarist reason in favour of present sacrifices (forgoing present welfare gains) to pursue additional future lives with positive welfare. But equally, there is no welfarist reason *against* present sacrifices to pursue additional future lives with positive welfare. Acting to reduce the risk of premature human extinction may therefore not have less merit, on welfare grounds, than acting to mitigate present suffering. That would be so, at least, if the choice problem looked somewhat like that described in Table 3.2.

4.3 Greedy neutrality and non-longtermism

We see that neutrality does not preclude choice conclusions that depend on the number of people who will live and the quality of their lives: great gains in worthwhile future lives may be worth (or can at least *offset*) sacrifices in welfare for those in the present. That should already give defenders of neutrality pause—this approach to welfare does not insulate present decision making from considerations of varying population size, at least not to the extent that one might have hoped. It gets worse, however. Neutrality does not preclude further disturbing implications that are not shared by totalist approaches. We will see that not only are the advantages of the neutral compared to the total approach to welfarist reasons less pronounced than first appearances suggest, but the former also has marked disadvantages.[11]

To see this, let us consider now the other side of greedy neutrality: cases in which welfare *gains* to the present are swallowed up by the addition of future lives with positive welfare. The recipe for generating such cases of swallowing up is similar to that outlined above: Take an 'original population'. This time add n lives with extremely high welfare; some finite level of welfare will work in the recipe since we are assuming that the neutral range is unbounded. Again, call this the 'augmented population'. It is, by construction, incommensurable with the original population with respect to the welfarist good. Now there will be various populations of the same size as the augmented population that are *worse* than that population. These '+n-populations' are thus *not better* than the original population; they are either incommensurable with or worse than the original population. But some of these +n-populations will involve welfare gains for the original population. All that is required for a population to qualify as a +n-population is that the sum welfare gains for the original population are outweighed by at least as much sum welfare loss for the added n people, compared to their welfare in the augmented population (i.e. extremely high welfare). These are the populations in which the added lives 'swallow up' gains to the original population.

This is a very troubling implication for neutrality, at least when formulated with the features specified above.[12] Note that the first feature is a neutral range that is unbounded from above, extending to positive infinity. It is arguably that first feature, rather than the second which specifies an average welfare approach to comparing populations of the same fixed size, which yields the most trouble. When the neutral range is unbounded from above, the added lives to the original population *could have had any positive welfare whatsoever* and the resulting augmented population would still count as incommensurable with the original population. So whatever the *actual welfare* of the added lives, a sum welfare loss of any size whatsoever is incurred, relative to a population that fits the recipe of the augmented population. This loss can outweigh any sum welfare gain to the original population. In short, neutrality does not preclude a very severe kind of swallowing up of gains to an

[11] This is not surprising in light of the 'impossibility theorems' of population ethics (see, for instance, Arrhenius n.d.). Standard forms of totalism notoriously imply the 'repugnant conclusion'. Alternatives designed to avoid that conclusion have other problems, for instance they imply the 'sadistic conclusion'. Neutrality is somewhere in the middle. It may lead to a *less* repugnant conclusion, but the other side of the coin is that it leads to a *less*, say, sadistic conclusion. In other words, we might expect that a neutral approach to the welfarist good will not have the extreme counter-intuitive properties of other accounts, but will have many counter-intuitive properties nonetheless.

[12] It amounts to a violation of an axiom known as *Dominance Addition* which Gustaf Arrhenius (n.d., 307) articulates as follows: 'An addition of lives with positive welfare and an increase in the welfare of the rest of the population doesn't make a population worse, other things being equal.'

original population. We see that adding *any number of lives to an original population*, at *any positive level of welfare*, 'swallows up' *any sized welfare gain* to the original population. That is, the resulting population is *no better* (incommensurable, or even worse) than the original population.

We can refer to this as a (worrying) *non-longtermist* implication of neutrality: increasing the size of the human population, even if this is a *win-win* scenario with no intertemporal trade-offs, is no better than the status quo. So for instance, the mitigation of climate change, or fantastic advances in medicine, insofar as they *increase the size of the human population by adding worthwhile lives*, are not better than the status quo, even if they also come with many welfare benefits for those in the present. We strangely have no positive welfarist reason under neutrality to pursue these win-win options.[13]

The way neutrality can swallow up gains in welfare is pertinent to Parfit's (1984, 453) pointed question (and yet is easily overlooked)[14]:

Compare three outcomes:

(1) Peace.
(2) A nuclear war that kills 99% of the world's existing population.
(3) A nuclear war that kills 100%.

Outcome (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences?

Parfit himself suggests that most people believe the greater difference is between (1) and (2). The neutral approach to welfare, unlike the totalist approach, seems to partially respect that intuition. Neutrality allows that (2) and (3) are incommensurable (or even that (2) is worse). We can assume that (2) is at least as good as (3) for all those bound to exist, and better for some of those people whose lives were not shortened so dramatically. But (2) also contains some number of extra people with worthwhile lives who may 'swallow up' the aforementioned gains (depending on the precise version of neutrality that is adopted). That would be the case if (2) were worse than *some* same-sized population that is incommensurable with (3) in that it only differs from (3) in containing extra lives. Then (3) would not be worse than (2). But at least in that case the gap in welfarist good between (2) and (3) would not be 'much greater' than the gap between (1) and (2). (Let us assume that the populations in (1) and (2) end up being the same size.)

However, that is not the whole story. For the same reasons that neutrality allows that (2) and (3) are incommensurable—reasons of 'greediness'—neutrality allows that (1) and (3) are incommensurable too. And yet the latter comparison is highly counter-intuitive. Consider the shift from (3) to (1): it involves large welfare gains for the small population who were bound to live—from the destitution and shortening of lives that precedes extinction to a state of world peace and increased longevity. And then there are all these extra

[13] This is presumably why Broome (2005, 410) claims that neutrality yields 'incredible' assessments of everyday choices: specifically, that doing nothing about climate change, or more locally, about road safety or taxation reform, is no worse than acting in a positive way to mitigate climate change, improve roads, or reform the taxation system, respectively.

[14] MacAskill (2022, 168) quotes this passage, and Setiya (2022) moreover appeals to Parfit's three options in his criticism of MacAskill's claims.

people living in the state of world peace, since humanity does not go extinct. Surely (1) is much better than (3). Again, the trouble is that, on neutrality, the additional persons on the world peace outcome may 'swallow up' the large gains in welfare to the 'original' people. After all, if the destitute people were joined by additional people who enjoyed not just world peace but super blissful world peace, this larger population outcome would have been incommensurable with the smaller destitute population. And plausibly, the outcome (1) is no better a population than the one with super blissed-out extra people: the downward shift in welfare of many from super blissful world peace to world peace outweighs the upward shift in welfare of relatively few from destitution to world peace.

5 Concluding remarks

We see that the neutral approach to welfare is more complicated than first impressions suggest. The greediness of neutrality means there is less of a contrast than might be hoped—when many additional lives are at stake—between the neutral and totalist approaches to welfare. On either approach, significant present sacrifices in welfare may be permissible (albeit, for neutrality, not required) if the result is a small reduction in the probability of some future threat, including threats whose severity turns on premature human extinction.

If Table 3.2 more or less accurately describes the choice problem we now face—whether we should aim to directly reduce the risk of some threat—neither the neutral nor the total approaches clearly furnishes us with strong reasons *not* to act on the future threat and instead focus on more proximate suffering. The extreme gains in welfare for the 'far-sighted' option under State 1 are relevant (and if large enough, swamp the comparison of options) for both kinds of approach.

One might yet hope that there is *some* plausible neutral approach to welfare that allows us to effectively sideline, in our welfarist reasoning, changes to the timing of human extinction or to population size more generally. After all, the analysis in section 4 depends on a specific version of the neutral approach. But it is far from assured that alternative versions of neutrality would change the overall story in ways that defenders of neutrality would welcome. For instance, the extent to which neutrality swallows up gains in welfare for an original population could be mitigated by lowering the upper bound of the neutral range to something not too much greater than zero.[15] But this would only lessen the difference between the neutral and totalist approaches to welfare. One might otherwise make different assumptions when it comes to comparing populations of fixed size. Instead of appealing to average welfare, one might rather compare fixed-size populations in terms of minimum welfare, for instance, or in some other way more sensitive to (in)equality. But many such differences will lead to changes in degree rather than kind when it comes to greedy neutrality. Similarly for the assumption of transitivity, I speculate. At any rate, the burden falls on defenders of neutrality to provide a fully worked out account that has clear advantages over the one presented in section 4.

That brings me back to the underlying aim of the chapter, which is to deflate the idea that the fundamental approach to welfare one adopts matters a great deal in assessing

[15] This would be what Rabinowicz (2009) calls a 'moderate' interpretation of the neutral range. His own critical-range utilitarianism involves a moderate neutral range—it is the range of critical values.

Table 3.3 Modified choice problem for a varying-sized population.

	State 1		State 2		...	State n	
	base	**spillover**	**base**	**spillover**	...	**base**	**spillover**
short-sighted	modest	none	modest	none	...	modest	none
medium-sighted	high	small	high	small	...	high	small
far-sighted	none	extreme+	extreme–	none	...	none	none

longtermist conclusions about future threats that affect the size of the population. The robustness of longtermist conclusions about what we have strong welfarist reason to do and not do hangs on whether the extreme empirical picture that is painted is in fact true. Do our choice problems really look like those described in Tables 3.1 and 3.2? If not, then we may *not* have strong welfarist reason to try to directly reduce the probability of premature human extinction, even on a simple totalist account of the welfarist good. We may rather have strong welfarist reason to pursue a different kind of project.

What would an alternative choice scenario look like? Table 3.3 offers just one example; it includes a 'medium-sighted' option. Perhaps this option involves institution building that is reasonably believed to be very good, over the long run, for the given number of people bound to exist, and to also enhance expected population size to some extent relative to the status quo. Moreover, the 'far-sighted' option in Table 3.3 has a different profile—it might possibly backfire, for instance, by making a long-enduring tyrannical regime more likely for the fixed population (an extremely bad outcome). In that case, even on a simple totalist account of the welfarist good, it is not obvious that there is welfarist reason to pursue the 'far-sighted' option. The 'medium-sighted' option might rather be optimal with respect to welfare.

I have said nothing here to help settle the question of whether Table 3.3 or Table 3.2 more accurately represents the choice problem we now face regarding future threats that affect population size. The point is rather that these are the kinds of hypotheses that we should be raising and scrutinising in assessing longtermist conclusions about what we have welfarist reason to do and not do. The totalist approach to welfare plays a relatively minor role in the 'longtermist's mathematics'.[16]

References

Arrhenius, G. (n.d.), *Population Ethics: The Challenge of Future Generations* (unpublished manuscript).
Broome, J. (2004), *Weighing Lives* (Oxford University Press).
Broome, J. (2005), 'Should We Value Population?', in *The Journal of Political Philosophy* 13/4: 399–413.
Frick, J. (2017), 'On the survival of humanity', in *Canadian Journal of Philosophy* 47/2–3: 344–367.

[16] Many thanks to Christian Barry, Hilary Greaves, and Alan Hájek for valuable comments on draft versions of this chapter. Thanks too to the seminar audience of the Institute for Futures Studies (IFFS), and especially Tim Campbell, for very helpful feedback. I am moreover grateful for support from the 'Climate Ethics and Future Generations' project at the IFFS, funded by Riksbankens Jubileumsfond (grant number M17-0372:1) and from an ANU Futures Scheme grant.

Frick, J. (2020), 'Conditional Reasons and the Procreation Asymmetry', in *Philosophical Perspectives* 34: 53–87.
Frick, J. (2022), 'Context Dependent Betterness and the Mere Addition Paradox', in J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan (eds.), *Ethics and Existence: The Legacy of Derek Parfit* (Oxford University Press), 232–263.
Greaves, H. and MacAskill, W. (this volume), 'The Case for Strong Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
MacAskill, W. (2022), *What We Owe the Future* (Basic Books).
Parfit, D. (1984), *Reasons and Persons* (Oxford University Press).
Nebel, J. M. (2019), 'Asymmetries in the Value of Existence', in *Philosophical Perspectives* 33: 126–145.
Rabinowicz, W. (2009), 'Broome and the Intuition of Neutrality', in *Philosophical Issues* 19: 389–411.
Rabinowicz, W. (2022), 'Getting Personal: The Intuition of Neutrality Reinterpreted', in G. Arrhenius, K. Bykvist, T. Campbell, and E. Finneron-Burns (eds.), *The Oxford Handbook of Population Ethics* (Oxford University Press), 114–141.
Setiya, K. (2022), 'The New Moral Mathematics', *Boston Review*, 15 August, 2022: https://bostonreview.net/articles/the-new-moral-mathematics/. Last accessed 11 January 2025.
Thomas, T. (2023), 'The Asymmetry, Uncertainty, and the Long Term', in *Philosophy and Phenomenological Research*, 107/2: 470–500.
Voorhoeve, A. (2014), 'How Should We Aggregate Competing Claims?', in *Ethics* 125/1: 64–87.

4

Prudential Longtermism

Johan E. Gustafsson and Petra Kosonen

According to

> *Longtermism*: Our acts' influence on the expected value of the world is mainly determined by their effects in the far future.[1]

Longtermism is counter-intuitive. It implies that our influence on the short term, which we normally focus on, is outstripped by our influence on the far future. Yet, given a total utilitarian view of expectations, there is a straightforward case for Longtermism. Even a small chance of a very large population living in the far future outweighs the importance of our acts' effects in the near future. Consequently, when evaluating acts, we can often simply ignore their short-term effects and focus on their effects in the far future.

It's less clear whether there is a similar case for Longtermism if we accept a person-affecting view, on which an outcome cannot be better than some other outcome unless it is better for someone.[2] Our acts today affect not only the number and quality of future lives but also who will exist in the future—so that the acts we can perform result in different people existing in the future due to the ripple effects of these acts.[3] So, if it can't be better or worse for someone to exist than not to exist, it seems that the only people we can make better off are those who already exist (and maybe people who will exist very soon).[4] In that case, if it's certain (or almost certain) that no one alive today will be alive in the far future, then person-affecting views lead to the rejection of Longtermism.[5]

But, in fact, there is a different path to Longtermism that is perfectly compatible with person-affecting views. Instead of total utilitarianism, this path appeals to

> *Prudential Longtermism*: Prudential Longtermism holds for person *S* if and only if our acts' overall influence on the expected prudential value for *S* is mainly determined by the effects of these acts in the far future.

[1] The term 'longtermism' was coined by MacAskill and Ord; see Ord (2020: 46, 306 fn. 27). And see Greaves and MacAskill (this volume) for a defence of (strong) longtermism.

[2] Narveson (1973: 80), Parfit (1984: 394–400), and Temkin (1987: 166–7).

[3] Parfit (1984: 351–5).

[4] It's controversial whether it can be better for a person to exist than not to exist. Williams (1973: 87), Parfit (1984: 487), and Broome (1993: 77) argue that it cannot, while Arrhenius and Rabinowicz (2015: 427–32) argue that it can.

[5] Bostrom (2003: 312). By contrast, asymmetrical views on which creating happy lives does not make the world better but creating unhappy lives does make the world worse, may lead to Longtermism. See Thomas (2023: 494–5). (See also Mogensen this volume and Steele this volume.)

Johan E. Gustafsson and Petra Kosonen, *Prudential Longtermism* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Johan E. Gustafsson and Petra Kosonen 2025.
DOI: 10.1093/9780191979972.003.0004

If Prudential Longtermism is false for all currently existing people, then normative views on which only these people matter lead to the rejection of Longtermism.[6] Or, at least, they do so if we assume (as seems plausible) that a current person's well-being can only be determined by effects in the far future if they affect the well-being of some individual in the far future for whom the current person is justified in having prudential concern.[7]

In this chapter, we will explore whether Prudential Longtermism is true. Prudential Longtermism depends mainly on the feasibility of different forms of life extension. But, as we shall see, it also depends on what relation matters in survival and on how we should aggregate personal value in cases of fission—that is, cases in which there are multiple individuals in the future who are all related to a person (as the person is now) in the way that matters for survival.

We may distinguish between different strengths of Prudential Longtermism:

> *Strong Prudential Longtermism*: Strong Prudential Longtermism holds for person *S* if and only if our acts' overall influence on the expected prudential value for *S* is *overwhelmingly* determined by their effects in the far future.[8]

> *Weak Prudential Longtermism*: Weak Prudential Longtermism holds for person *S* if and only if our acts' overall influence on the expected prudential value for *S* is *mostly* determined by their effects in the far future.

If Weak Prudential Longtermism holds for someone, then the far future matters more in expectation than the near future for their prudential value. By contrast, if Strong Prudential Longtermism holds for someone, then the far future matters overwhelmingly more than the near future for their prudential value, and, for their prudential concerns, we can often simply ignore our acts' short-term effects and focus on the long-term effects.

We will discuss whether Weak and Strong Prudential Longtermism hold for some currently existing people and whether this means that even person-affecting views lead to (impersonal) Longtermism. It's clear that there are things we could do such that we would have no hope of any prudential value after the short term. So, in our discussion, we will look for acts and technologies that may provide a lot of prudential value in the long term. By performing such acts rather than the acts that offer no expectation of long-term prudential value for some person, our acts have an enormous influence on their expected prudential value. And then Prudential Longtermism holds for that person.

As mentioned, the case for Prudential Longtermism relies on the feasibility of extreme life extension.[9] There are a number of ways in which we might be able to extend our healthy

[6] Many of the new insights from the recent flurry of research on effective altruism have yet to be applied to prudential concerns—an endeavour which we may call *effective prudentialism*. If we're effective when it comes to spending 10% of our income on altruistic causes, why be careless with the remaining 90%?

[7] Scheffler (2013: 73; 2018: 44) claims that, for many people, the value of their current activities depends on there being future generations continuing these activities—even though they accept that they, themselves, will die young. If so, Prudential Longtermism might be true for them, since their current well-being depends on the existence of other people in the far future. Scheffler (2018: 53–7), however, denies that the existence of future generations after our deaths would provide us with prudential reasons—see also Greaves (2019: 138–40) for some objections to Scheffler on this point. In this chapter, we will not explore this posthumous route to Prudential Longtermism.

[8] MacAskill (2019) defines Strong Longtermism as the view on which we should be most concerned about the long-run outcomes, while Very Strong Longtermism is defined as the view on which the long-term outcomes are of overwhelming importance.

[9] Bostrom (2003: 312).

lifespans. While these forms of life extension may be far-fetched, we will argue that, for some of them, even a small chance of them working is sufficient to support Prudential Longtermism. Hence, while we defend Prudential Longtermism, we are not claiming that any of these forms of life extension are likely to work.

1 Anti-ageing

Anti-ageing is the attempt to stop, or even reverse, ageing.[10] Research on anti-ageing has made some progress.[11] Could anti-ageing, by itself, lead to Prudential Longtermism? If it succeeds in stopping or reversing ageing, it could, of course, significantly lengthen our lives. But, even if we stop ageing, we may still die from other causes. Given a 0.13% chance of death per year (the proportion of people aged 30–31 who died in the U.S. in 2019), one has a 27% chance of surviving for 1,000 years and just a 0.00022% chance of surviving for 10,000 years. And one's life expectancy is $1/0.0013 \approx 770$ years.[12] This estimate assumes that the annual background risk of death (from injury or illness) won't change, and it doesn't take into account rare events, such as wars, global catastrophes, or existential risks.[13] Is 770 years a sufficiently long life expectancy to lead to Prudential Longtermism?

Let the next 100 years constitute the short term, and let the long term start thereafter.[14] And let us assume (somewhat arbitrarily) that a technology leads to Strong Prudential Longtermism if and only if, due to this technology, a person's expected number of life years in the long term (the period starting after the next 100 years) is at least 10,000 times as great as their expected number of life years in the short term (the next 100 years). This will be true if their life expectancy is at least 1 million (plus 100) years, assuming that the person will certainly live for 100 years.[15] If there were such a technology, then it would be plausible that our acts' overall influence on the expected well-being of some currently existing person is overwhelmingly determined by our acts' effects in the far future. Next, let us assume that a technology leads to Weak Prudential Longtermism if and only if a person's expected number of life years in the long term (the period starting after the next 100 years) is greater than that person's expected number of life years in the short term (the next 100 years).

[10] See de Grey and Rae (2007) for an overview and defence of the feasibility and prudential value of anti-ageing, and see Bostrom (2005) and Bostrom and Ord (2006: 676–7) for further defences of the desirability of lengthening our healthspans.

[11] For an optimistic overview of recent advances in anti-ageing research, see Partridge, Fuentealba, and Kennedy (2020). Ramakrishnan (2024: 203–20) and Crimmins (2015: 908–9) offer a less optimistic take, the latter claiming that the necessary interventions may need to be done at a very young age. So, even if anti-ageing would be invented in our lifetimes, it may be too late for current adults. In other words: if you can read this, it may be too late for you.

[12] Based on data from Arias and Xu (2022: 10). Similarly, Bostrom and Roache (2007: 124) estimate that, if we lived at the mortality rate of someone in their late teens or early twenties, our life expectancy would be around 1,000 years.

[13] Ord (2020: 167) estimates that the risk of human extinction within the next 100 years is already 1/6.

[14] In Greaves and MacAskill (this volume), the far future is 'everything from some time t onwards, where t is a surprisingly long time from the point of decision (say, 100 years)'.

[15] Temkin (2008: 202–4) argues that an extremely long life may get boring after a while, noting that he has listened to his favourite music (mostly late 60s and early 70s rock) so much that it no longer gives him much pleasure. For a similar complaint, see Williams (1973: 90). Note that, if our lives would inevitably become more or less neutral in the far future, then we would not make a difference to how much well-being we would have in the far future even if we develop anti-ageing technology. So might we, like Temkin, eventually run out of new pleasures? That seems unlikely. There is a simple solution to Temkin's predicament: Try some new music.

Then, assuming that the long term does not provide opportunities for far greater or far lower welfare per unit of time than the short term, it's likely that our acts' overall influence on some currently existing person's expected well-being is mostly determined by their effects in the far future.

How high must our credence in anti-ageing working be in order for it to lead to Weak Prudential Longtermism? By anti-ageing working for a person, we mean anti-ageing being successfully used by them. With p being the constant probability of death each year if anti-ageing works (which we have assumed to be 0.13%), we find that the expected years of life in the short term (that is, the next 100 years) if anti-ageing works is

$$\sum_{n=1}^{100}(1-p)^n \approx 93.7.$$

Let q be the probability of anti-ageing working. And assume that a person's current life expectancy without any new life-extension technology is 50 years (the U.S. life expectancy at age 30 in 2019).[16] Now, anti-ageing alone leads to Weak Prudential Longtermism if

$$\left(\frac{1}{p}-93.7\right)q > 93.7q + 50(1-q).$$

On the left side of the equation in the brackets we have the expected number of life years in the long term (after the next 100 years) conditional on anti-ageing working. This is then multiplied by q, the probability of anti-ageing working, to get the overall expected number of life years in the long term. (Note that the expected number of life years in the long term if anti-ageing does not work is assumed to be zero.) On the right side of this equation, we first have the expected number of life years in the near term conditional on anti-ageing working (93.7 years), multiplied by the probability of anti-ageing working. Next we have the expected number of life years in the near term conditional on anti-ageing not working (50 years), multiplied by the probability of anti-ageing not working. Adding these two together gives us the overall expected number of life years in the near term. If the left side of the equation is greater than its right side, then the expected number of life years in the long term is greater than the expected number of life years in the near term.

Hence anti-ageing leads to Weak Prudential Longtermism if q, the probability of anti-ageing working, is greater than 8%. Then, the person's expected number of life years in the long term is greater than their expected number of life years in the short term.[17]

But anti-ageing alone does not lead to Strong Prudential Longtermism (as we have defined it). Even assuming that anti-ageing is guaranteed to work, the expected number of life years in the long term is less than eight times greater than the expected number of life years in the short term, given a 0.13% yearly chance of death. Of course, we may be able to decrease our yearly risk of death and thereby improve our chances of survival significantly.

[16] Using data from Arias and Xu (2022: 3).

[17] But anti-ageing alone need not give us Weak Prudential Longtermism if we accept the Multiplicative View of Continuity Strength (discussed in section 3) and the weight of Relation *R* holding between consecutive person-slices is less than one.

In order to get 10,000 times as great an expectation of number of life years in the long term as in the short term, we need the annual risk of death to be at most one in a million. But this, of course, assumes that anti-ageing works. Since there is significant uncertainty about the feasibility of anti-ageing, the annual risk of death needs to be even lower in order for anti-ageing alone to lead to Strong Prudential Longtermism.

2 Cryonics

Cryonics is the process of storing a person's brain (or whole body) at very low temperature after their legal death in the hope that they will one day be revived. A current way to cryopreserve a brain is through vitrification, which hardens water like glass without crystal formation that would cause damage to cells. The brain is then kept cool with liquid nitrogen. The hope is that this process will preserve the brain without further tissue degradation and that medical science will eventually make advances that allow the stored brains to be revived (and repaired) back to a healthy life.[18]

One worry about cryonics is whether it can preserve memories.[19] Many philosophers believe that psychological continuity is what matters in survival.[20] On this view, an outcome is as bad as death for a person unless they are psychologically continuous with someone in that outcome. Psychological continuity in turn consists in overlapping sequences of psychological connections. And these connections are usually taken to be memory relations, that is, the relation of the current person's experiences being remembered by the future person.[21] So, on these views, cryonics does not preserve what matters in survival if it fails to preserve one's memories.

Yet there are other candidates for what matters in survival. Some believe that *physical* continuity is what matters.[22] On these views, an outcome is as bad as death for a person unless that person has the same brain (or enough of the same brain) as someone in that outcome. Thus, on these views, cryonics could preserve what matters in survival even if it fails to preserve memories (or any other psychological connections)—as long as it's possible to revive the same spatio-temporally continuous brain.

Does cryonics in combination with the technology to revive a cryopreserved brain lead to Strong Prudential Longtermism? Even if cryonics combined with such a technology leads to a successful revival, it is still open to worries about fatal injuries that permanently destroy the brain after the revival. So, even if it's possible to revive the spatio-temporally continuous brain after cryopreservation and this brain could be given a new body, that brain may still be damaged beyond the possibility of revival. The annual risk of brain destruction (during those years in which the brain is not cryopreserved) would have to be at

[18] Merkle (1992: 6; 1994: 16). The feasibility of cryonics is controversial. While some scientists think that it could work—see, for example, Benford (2005)—others claim that it won't—see, for example, Miller (2004) and Ramakrishnan (2024: 199–203). For a defence of the practice of cryonics, see Shaw (2009).

[19] Doyle (2018: 124). Vita-More and Barranco (2015: 458), however, claim to have made progress in preserving long-term memory in worms after cryopreservation.

[20] According to Minerva (2018, 10), the dominant view among supporters of cryonics is that a person is fundamentally the information stored in a brain.

[21] Parfit (1984: 205), however, suggests that other psychological relations may also matter. Plausibly though, if memory connectedness does not hold, this is likely to be accompanied by the rupture of other connections.

[22] Nagel (1986: 40) and Unger (1990: 109).

most one in a million in order to get at least 1 million life years in expectation. This bar for the risk of brain destruction is still too low for Strong Prudential Longtermism to be true for anyone. Hence cryonics in combination with anti-ageing and the technology to revive a cryopreserved brain at most gives us Weak Prudential Longtermism. Still, cryonics (like anti-ageing) might buy us time for finding better ways of extending life.

3 Uploading

Uploading (also known as whole-brain emulation) is the process of scanning a person's brain and loading the information onto a computer, where the brain is then simulated.[23]

A standard worry about uploading is whether the simulation will be conscious.[24] A zombie simulation would not (assuming hedonism) have any well-being, so it would be prudentially worthless, intrinsically. Another worry is whether the upload would be identical to the current person, that is, whether the current person would be identical to their simulated self.[25] A more pressing worry, however, is whether the current person would stand in the relation that matters in survival to their simulation. The views of personal identity on which one could plausibly be identical to one's simulation are reductionist views where personal identity just consists in an impersonal mental relation holding uniquely.[26] Here, an impersonal relation is a relation that can be completely described without mentioning people. But, if personal identity can be reduced to an impersonal relation (such as psychological continuity) holding uniquely, it seems that we should also care about this relation when it holds from one to many rather than only in the case when it holds from one to one.[27]

The most influential reductionist view is that psychological continuity is what matters in survival. As mentioned, psychological continuity is the holding of overlapping sequences of psychological connectedness. Psychological connectedness is a direct psychological connection between a person at one time and a person at another time, such as the person at the latter time remembering (or quasi-remembering) the experiences of the person at the earlier time.[28]

To discuss psychological continuity and the relation of what matters in survival, we will adopt a perdurance framework in which we analyse persistence in terms of person-slices, that is, instantaneous temporal parts of people.[29] (But our discussion could also be done in an endurance framework, changing what needs to be changed.) We will represent psychological connectedness by Relation *C*, defined as follows:

[23] See Sandberg and Bostrom (2008: 7–15) for an overview of uploading.
[24] Chalmers (2010: 44–8).
[25] Aaronson (2016: 210–1) and Chalmers (2010: 48–63).
[26] Parfit (1984: 263). Actually, the structure of the account must be somewhat more complicated. See Gustafsson (2019: 2314–5).
[27] Parfit (1971: 4–14; 1984: 256, 261–4) and Gustafsson (2018: 745–50).
[28] Quasi-remembering is just like remembering except that the remembered person needn't be the same as the remembering person. See Shoemaker (1970: 271).
[29] Brink (1992: 215–6). To avoid overlap between stages, we rely on person-slices rather than person-stages, which need not be instantaneous. For person-stages, see Perry (1972: 467) and Lewis (1986: 202). We may wish to allow that person-slices, rather than being instantaneous, have the minimal duration necessary to be able to have well-being. But, if so, we need to individuate the slices in a way that avoids overlap between slices.

> Person-slice x is C-related to person-slice y (xCy) $=_{df}$ x is psychologically connected to y with the right kind of cause and x is either simultaneous with y or earlier than y.

We will represent psychological continuity by Relation R, which we define in terms of the holding of overlapping chains of C-relations:[30]

> Person-slice x is R-related to person-slice $y =_{df}$ either xCy or yCx, or there are person-slices $z_1, z_2, \ldots, z_n$ such that either
> (i) $xCz_1, z_1Cz_2, \ldots, z_{n-1}Cz_n, z_nCy$ or
> (ii) $yCz_1, z_1Cz_2, \ldots, z_{n-1}Cz_n, z_nCx$.

Both Derek Parfit and David Lewis wobble a bit on whether to put the emphasis on continuity or on connectedness—that is, whether it's Relation C or Relation R that matters (or both).[31] The distinction is crucial for determining the amount of prudential value someone gets from uploading.

3.1 A very long simulation

One reason to think that uploading may lead to Prudential Longtermism is that the uploads can live on for a very long time.[32] If the simulations of a current person S gradually change their psychology over time, they may eventually stop being C-related to S as S is now, even though they would still be R-related to S.[33] Since prudential concern is plausibly forward (rather than backward) looking, the simulations need not have any special interest in continuing to be directly psychologically connected to S.[34] So we may suspect that they will gradually let go of their memories of S in order to make room (in computer memory) for more useful knowledge.[35] Hence, if Relation C is what matters, it seems that uploading would not lead to Prudential Longtermism in virtue of a very long-lasting simulation. But, if Relation R is what matters in survival, it seems that, as long as the simulation is kept running, one's relation to one's simulation contains what matters. And, if civilization survives and people have some interest in keeping the simulation running, then the simulation may run for a very long time.

Assuming that S, as S is now, is R-related to a large number of person-slices of a long-lasting simulation, how much prudential value does this provide for S? That depends on three factors: (i) how much S's relation to each of these person-slices matters, (ii) how

[30] McMahan (2002: 50).
[31] Parfit (1971: 21; 1984: 262) and Lewis (1976: 18).
[32] Dyson (1979: 456) suggests that a finite amount of physical energy could be used to simulate an infinite amount of subjective time.
[33] Lewis (1976: 29–31).
[34] Parfit (1984: 174–7), however, challenges this bias towards the future.
[35] But, if the simulations accept evidential decision theory, they may wish to keep memories of their earlier person-slices, because letting go of those memories would be evidence that the later person-slices will also choose to let go of their memories. For evidential decision theory, see Jeffrey (1965: 1–6; 1983: 1–6), Gibbard and Harper (1978: 129), and Ahmed (2014: 43–6). Alternatively, could one try to cultivate a false belief in backward-looking prudential concern? It seems that, if one can tell that a philosophical view is implausible, then these descendant simulations of us would be able to do so too.

well-off these person-slices are, and (iii) how the well-being of these person-slices should be aggregated.

Let a *life-path* be a maximal aggregate of person-slices that are related by what matters to each other, that is, an aggregate of person-slices such that (i) each slice in the aggregate is related by what matters to all slices in the aggregate and (ii) no person-slice that is not in the aggregate is related by what matters to all slices in the aggregate. The important thing about life-paths is that they are unified in the sense that the relation that matters does not branch, as all person-slices in a life-path are related by what matters to all others in that life-path. (On some views of personal identity, a life-path is a person.[36] But we do not need to assume this.)

Now, regarding the aggregation of the well-being of the future person-slices, consider

> *The Single Life-Path Total View*: Within a single life-path, the overall prudential value of a risk-free prospect for person *S* now is the sum total, for all future person-slices within that life-path, of the well-being of that slice multiplied by the weight of the *R*-relation between that slice and *S* (as *S* is now).

On this view, a person's future momentary well-being is added up, in proportion to the weights of the *R*-relations, to get the prudential value of their future. The Single Life-Path Total View implies

> *The Single Life-Path Repugnant Conclusion*: For any possible life-path in which each person-slice has very high well-being and is *R*-related to person *S* (as *S* is now), there is a possible life-path that is better for *S* even though each of its person-slices has barely positive well-being.

This conclusion implies that, for any number of years that a person could live at a high momentary well-being level, there is some number of years, during which they have barely positive momentary well-being, that is better for them.[37] The Single Life-Path Total View also entails the following variant where the weights of the *R*-relations are different but well-being is held constant:

> *The Weighted Single Life-Path Repugnant Conclusion*: For any possible life-path in which each person-slice has positive well-being and is strongly *R*-related to person *S* (as *S* is now), there is a possible life-path that is better for *S* even though each of its future person-slices is barely *R*-related to *S* (as *S* is now) and, in both cases, the person-slices within those life-paths have the same positive well-being.

While these conclusions may seem counter-intuitive, they seem less so than the corresponding Repugnant Conclusion in population ethics.[38]

Moreover, we can defend these single life-path conclusions with a mere-addition argument:[39] Adding a long life that is at each point minimally positive in well-being to a

[36] Our definition of a life-path corresponds to Lewis's (1976: 22) definition of a continuant person.
[37] See McTaggart (1927: 452–3), Parfit (1986: 160), Crisp (1997: 24–5), and Temkin (2012: 119).
[38] Parfit (1984: 388).
[39] This argument is analogous to the Mere-Addition Paradox in Parfit (1984: 419–41).

person's lifespan seems to be at least as good for them as their life without that addition. Then, making that person's life equal in quality throughout while increasing the average momentary level of well-being slightly seems to be better for them. Then, by the transitivity of *at least as good as*, we find that the end result—that is, a life that is at all times barely worth living—would be better for the person than their current lifespan (no matter how good their current lifespan is).[40]

But it's less obvious how to weigh the importance of being *R*-related to a person-slice. Relation *C* has a straightforward weighting: the proportion of how much of the earlier person-slice's psychological state the later person-slice shares or remembers. Since Relation *R* holds in virtue of overlapping sequences of *C*-related person-slices, it's compelling to adopt following view:[41]

> *The Multiplicative View of Continuity Strength*: Let a *weight-product* of a sequence of *C*-related person-slices be equal to the product of the weights for each *C*-relation in the sequence. The weight of Relation *R* holding between person-slices x and y is equal to the maximum weight-product of any sequence xCy or yCx or a sequence via person-slices z_1, $z_2, \ldots, z_n$ such that either
> (i) $xCz_1, z_1Cz_2, \ldots, z_{n-1}Cz_n, z_nCy$ or
> (ii) $yCz_1, z_1Cz_2, \ldots, z_{n-1}Cz_n, z_nCx$.

Note that, between two person-slices, there might be lots of sequences of overlapping *C*-relations and that the sequence with the greatest weight-product need not be the longest—it may even be the sequence consisting of a single direct *C*-relation between the two person-slices.

Does this view lead to Strong Prudential Longtermism given a successful upload with a long-lasting simulation? The trouble is that, once we allow the *C*-relations between the simulated person-slices to have weights that are less than 100%, the sum of the weights of the *R*-relation for all person-slices will converge relatively quickly, assuming that current

[40] If one is tempted to resist the Single Life-Path Total View, one could adopt

> *The Single Life-Path Average View*: Within a single life-path, the overall prudential value of a risk-free prospect for person *S* now is the sum total, for all future person-slices within that life-path, of the well-being of that slice multiplied by the weight of the *R*-relation between *S* (as *S* is now) and that slice divided by the sum total of all the *R*-relations weights. (Note that all person-slices are assumed to last equally long.)

On this view, a person's future momentary well-being levels are averaged over (while taking into account the weights of the *R*-relations) to get the prudential value of their future. This view, however, implies

> *The Single Life-Path Masochistic Conclusion*: It can be better for person *S* if, within a single life-path, *S* (as *S* is now) were related by what matters in survival to a small number of additional person-slices with negative well-being than if *S* (as *S* is now) were related by what matters in survival to a large number of additional person-slices with positive well-being (other things being equal).

This conclusion follows, because the person's average momentary well-being might be decreased less by the addition of the person-slices with negative well-being than by the addition of the person-slices with positive well-being. Another problem for the Single Life-Path Average View is that, if it is future-oriented, the prudential value of an immediate death is undefined as there would not be any future person-slices whose well-being can be averaged over. But, if we take the average over one's lifetime instead, then we get an analogous problem to the Egyptology objection to average utilitarianism: what happened in someone's distant childhood matters for which future is best for them. See McMahan (1981: 115) and Parfit (1984: 420).

[41] This view entails McMahan's (2002: 50) view that prudential concern is transitive: If the relation that matters holds to some extent between person-slice *x* and person-slice *y* and to some extent between *y* and person-slice *z*, then it holds to some extent between *x* and *z*.

memories (or whatever the *R*-relation consists of) are not retained at a higher rate than future memories. Each person-slice has prudential reasons to prefer being remembered by the next person-slice, so they would not opt to be forgotten by their immediate successor. But it seems that person-slices need not have any prudential reason to prefer that their predecessors are remembered by the next person-slice. So it seems that person-slices may opt to forget earlier person-slices in order to free up resources for more important information (or additional simulations). Let us therefore assume, to make the calculation simple, that person-slices only remember their immediate predecessor person-slice. Let each person-slice of the simulation be a year long (rather than instantaneous). And suppose that the well-being of each person-slice is constant at u. Let the weight of each *C*-relation be w. Then, given the Multiplicative View of Continuity Strength, the prudential value of an x-years-long simulation is

$$\sum_{i=1}^{x} uw^i = \frac{uw(w^x - 1)}{w - 1}$$

As the simulation lasts longer, this converges to

$$\sum_{i=1}^{\infty} uw^i = -\frac{uw}{w - 1}$$

To see that this does not favour Strong Prudential Longtermism, note that (given a positive well-being u and given that the weight w for the *C*-relations is positive and not greater than 100%) the infinite number of years after the next 100 years do not contribute 10,000 times more to the prudential value of the future than the next 100 years unless 99.9999% of each person-slice's psychology is retained each year.[42]

Would it be in each person-slice's interest that the next person-slice of the simulation remembers them to this extreme extent? It may seem that it would, because the more the next person-slice remembers them, the more the next slice (and the future) matters to them. But, if each slice needs to remember the last one completely, it seems that the simulation would constantly need more memory in order to store new knowledge. (Computational resources could also be used to create more simulations.) So it would make sense at some point to forget the last person-slice to some extent. But, if so, a long-lasting simulation does not (by itself) lead to Strong Prudential Longtermism.

Another potential way in which a long-lasting simulation may lead to Strong Prudential Longtermism is if the well-being levels of the person-slices of the simulation gradually get better. Even if the sum of the weights for the *R*-relations converges, it might still be that the overall prudential value increases faster and faster. With the addition of technological

[42] This results in a form of discounting of the future. But it is not a pure-time preference of the kind Sidgwick (1907: 381), Ramsey (1928: 543–4), Rawls (1971: 293; 1999: 259), and Parfit (1984: 125–6) object to. Yet Ahmed (2020) does object to this kind of psychological discounting. It's unclear, however, why we should accept his (2020: 247) Stationarity assumption that one takes at all times the same attitude towards well-being at the same distance in the future. If, on Monday, one knows that one will lose a lot of memories on Thursday and lose very few memories before then, then one plausibly cares more on Monday about one's Wednesday well-being than one will care on Wednesday about one's Friday well-being.

advances over time, we may be able to achieve increasingly higher welfare, and this might offset the decreasing weights of the *R*-relations to these distant person-slices.

3.2 Branching simulations

Earlier, we distinguished the view that some relation matters in survival from the view that personal identity matters in survival—even if personal identity only consists in the former relation holding uniquely (that is, without branching). When we assess the prudential value of uploading, the difference between these views matters a great deal. The reason it matters is that, once we have created a simulation of someone's brain, we can create many more.[43]

If we allow for branching in the relation that matters, we allow that someone can stand in the relation that matters to two (or more) simultaneous person-slices (which do not stand in the relation that matters to each other). But how should we aggregate the well-being of future person-slices in branching cases—that is, cases of fission?[44]

Suppose that a person *S* will undergo uploading and that either (*A*) one simulation will be created and it will enjoy four years of high momentary well-being or (*B*) that simulation and a separate simulation will be created and each of these simulations will enjoy three years of high momentary well-being at the same momentary well-being level as in *A* (Ω denotes non-existence):

	S_1	S_2
A	4	Ω
B	3	3

Consider next, expanding the additive approach of the Single Life-Path Total View to cases involving multiple life-paths,

> *The Prudential Total View*: The prudential value of a risk-free prospect for person *S* is equal to the sum total of the well-being of every person-slice that *S* (as *S* is now) is related to by the relation that matters, where the well-being of each slice is weighted by the strength of that relation.[45]

On this view, *S* would be better off if two three-year simulations were created instead of one four-year simulation, that is, *B* is prudentially better than *A*.

[43] See Dainton (2012: 56) for a discussion of fission through multiple uploads.

[44] One benefit of fission is that it allows one to become multi-planetary in the sense that one could stand in the relation that matters in survival both to future people on Earth and simultaneous (or space-like separated) future people on Mars. This allows one to survive a catastrophe that eliminates all life on one of these planets. This is, of course, analogous to the quest to safeguard humanity as a whole by becoming multi-planetary. See Sagan (1994: 371), Parfit (2017: 436), and Ord (2020: 392–3 fn.16), but compare Ord (2020: 194).

[45] Holtug (2001: 55; 2010: 118) presents a person-focused (rather than person-slice-focused) prudential total view. And Ross (2014) argues against a similar view.

Let *a person's life-paths* be the life-paths that have that person's current person-slice as a member. The Prudential Total View entails the following conclusion:[46]

> *The Prudential Repugnant Conclusion*: For any outcome in which each of person *S*'s life-paths has a great prudential value for *S*, there is an outcome that is better for *S* even though each of *S*'s life-paths in that outcome has a barely positive prudential value for *S* (and, in both outcomes, the person-slices within the life-paths have the same weights for the *R*-relations).

In the case of uploading, this conclusion implies that, for any number of simulations of *S* with very high well-being, there is a prudentially better outcome for *S* that contains a much larger number of simulations of *S*, each with a barely positive well-being level (while holding the weights of the *R*-relations fixed).

The Prudential Total View also entails the following variant, where the weights of the *R*-relations are different (but the well-being contained in each life-path is held constant):[47]

> *The Weighted Prudential Repugnant Conclusion*: For any outcome in which all the future person-slices of person *S*'s life-paths are strongly *R*-related to *S* (as *S* is now), there is an outcome that is better for *S* even though all the future person-slices of *S*'s life-paths in that outcome are barely *R*-related to *S* (as *S* is now) and the sum total of well-being of person-slices in each life-path is the same in both outcomes.

In the case of uploading, this conclusion implies that, for any number of simulations that are at all times strongly *R*-related to *S* as *S* is now, there is a prudentially better outcome that contains a much larger number of simulations that are barely *R*-related to *S* as *S* is now (holding the well-being of the simulations fixed).

We can contrast the Prudential Total View with an average view. The latter is slightly more complicated than one might think, since we still would like to maintain a sum-total view concerning the aggregation of momentary well-being over time (within one life when there is no fission).[48] To do so, we will introduce some terminology. As before, let a life-path be a maximal aggregate of *R*-related person-slices—that is, an aggregate of person-slices such that (i) each slice in the aggregate is *R*-related to all slices in the aggregate and (ii) no person-slice that is not in the aggregate is *R*-related to all slices in the aggregate.[49] Let a *successor* to a person-slice *x* be the person-slice that is next after *x* in a life-path of which *x* is part. Let a *fission-slice* be a person-slice with multiple successors.

> *The Prudential Average View*: Evaluate the prudential value of each life-path by the Single Life-Path Total View. Assume that fission-slices are followed by a chance node with an equal probability of being followed by each of that slice's successors. Hence we transform

[46] Gustafsson and Kosonen (2024: 1924 fn. 10). We assume here that well-being can be represented by a real-valued function.

[47] See Holtug (2001: 60).

[48] See fn. 40 for an argument against the average view concerning the aggregation of momentary well-being over time.

[49] Lewis (1976: 22).

prospects with fission into prospects of uncertainty. The prudential value of a prospect for person *S* is equal to *S*'s expected well-being in the transformed prospect.[50]

On this view, we treat the prospect of the two three-year simulations as if it were a 50-50 lottery between each of the two simulations being implemented on its own without the other. Hence, on the Prudential Average View, the prudential value of the two three-year simulations is the same as the prospect of a single three-year simulation—which is worse than the single four-year simulation.

Which of these answers is more plausible? Combining Parfit's Division Argument and his Mere-Addition Paradox, there is a straightforward argument for the answer of the Prudential Total View.[51] Consider, in addition to *A* and *B*, a third prospect A^+ that is just like *A* except that a second simulation is also implemented and this additional simulation has the same momentary well-being level as the first simulation but is only run for one year:

	S_1	S_2
A	4	Ω
A^+	4	1
B	3	3

It seems that, if simulation S_1 in *A* provides what matters in survival, then the same simulation in A^+ should also provide what matters in survival. The only difference in A^+ is that, in addition to S_1, there is another simulation to which *S* also stands in the relation that matters. So, in terms of what matters in survival, A^+ should be at least as great a success as *A*.[52] Consequently, A^+ should be at least as good as *A* for *S*. (Hence we should reject the Prudential Average View.[53]) Next, compare A^+ and *B*. Prospect *B* differs from A^+ in that S_1 lives for one year less but S_2 lives for two more years. Given that *S* stands in the relation that matters to *both* simulations, in terms of prudential value, the two extra years for S_2 in *B* should outweigh the single extra year for S_1 in A^+. So *B* is better than A^+ for *S*. Then, by the transitivity of *at least as good as*, we find that *B* is better than *A* for *S*.[54] Changing what needs to be changed, this argument also shows that we should accept the Prudential Repugnant Conclusion and the Weighted Prudential Repugnant Conclusion.[55]

[50] Tappenden (2011: 302). Another way to formulate the Prudential Average View would be to average over the well-being of all life-paths. This, however, would imply that, if person *S* first splits into two and much later one of the fission products splits multiple times (while the other does not), then that fission product's well-being (even before the later splits) would have overwhelmingly more influence on *S*'s prudential value, because it is part of multiple life-paths. Thus this results in a form of double counting of well-being.

[51] Parfit (1984: 419–26).

[52] Parfit (1971: 5; 1984: 256, 261–2; 1993: 24–5; 1995: 42).

[53] This argument is adapted from Gustafsson and Kosonen (2024: 1923–5).

[54] The transitivity of *at least as good as* can be taken to be an analytic principle of logic. See Broome (2004: 50–63). Or it can be defended with a money-pump argument. See Gustafsson (2022a: 39–44).

[55] The Prudential Average View also, implausibly, entails

The Masochistic Conclusion: It can be better for person *S* if *S* were to get some number of additional life-paths with negative prudential value than if *S* were to get some number of life-paths with positive prudential value (other things being equal).

This is a one-person counterpart to the Sadistic Conclusion. See Arrhenius (2000: 54). See also fn. 40 for a life-path version. To see how the Masochistic Conclusion follows, consider prospects *A* and *B*. In prospect *A*, there are three

Given that we adopt the Prudential Total View, rather than the Prudential Average View, we seem to have a route to Strong Prudential Longtermism. If we create not just one simulation of some currently existing person *S* but a large number of simulations, *S*'s prudential value from these simulations increases in proportion to the number of simulations. Moreover, each one of these simulations is in much the same situation, as they also increase their prudential value from the future the more simulations there will be of them. And, in turn, these further simulations are in much the same situation, as they can increase their prudential value by creating even more descendant simulations. Hence it seems that we would get an explosion of more and more simulations that are *R*-related to *S* as *S* is now.[56] Since this increase in the number of simulations will outweigh the diminishing weight of the *R*-relation between *S*, as *S* is now, and the simulations as they get more distant from *S*, *S* will (at least in expectation) get most of their prudential value from this enormous number of simulations in the far future. Hence, if we have a sufficiently high credence in uploading working and *S* having sufficiently many future branches (and in us being able to non-negligibly affect the welfare levels of those branches), then we get Strong Prudential Longtermism.

Given this explosion in the number of simulations, there will be a similar explosion in the demand for computational resources. This would put everyone in competition with everyone else for any available computational resources. Could this competition be avoided? It seems that it could. If the relation that matters in survival can split into multiple branches, it seems that it could also merge from many branches into one.[57] In the case of Relation *R*, this would happen when a person-slice is psychologically connected (that is,

separate simulations of person *S*: S_1 has a well-being of 13, whereas S_2 and S_3 have a well-being of 1. In prospect *B*, there are just two simulations of *S*: S_1 has a well-being of 13 (just like in *A*) and S_2 has a well-being of –1:

	S_1	S_2	S_3
A	13	1	1
B	13	–1	Ω

Here, the Prudential Average View entails that, for *S*, the prudential value of *A* is (13 + 1 + 1)/3 = 5 and the prudential value of *B* is (13 + (–1))/2 = 6. Thus it entails that *B* is better than *A* for *S*—which is an instance of the Masochistic Conclusion.

[56] Furthermore, note that, once a scan has been made of a person *S*, any replicas created from that scan no longer stand in the relation that matters in survival to *S* as *S* is after that scan. Or, at least, the replicas from the old scan wouldn't do so if, as seems plausible, the relation that matters in survival is temporally ordered (like Relation *R* is defined in this chapter). Some people take the relation that matters to be temporally unordered. For example, we could define a temporally unordered variant of psychological continuity as follows:

> Person-slice *x* is *C*′-related to person-slice *y* ($xC'y$) $=_{df}$ *x* is strongly psychologically connected to *y* with the right kind of cause.
>
> Person-slice *x* is *R*′-related to person-slice $y =_{df} xC'y$ or there are person-slices $z_1, z_2, \ldots, z_n$ such that $xC'z_1$, $z_1C'z_2, \ldots, z_{n-1}C'z_n, z_nC'y$.

The trouble is that, in a standard fission case where Wholly splits into Lefty and Righty, it seems plausible that Wholly is *C*′-related to each of Lefty and Righty. But it does not seem plausible that Lefty has what matters in survival to Righty, even though Lefty is *R*′-related to Righty. (See Gustafsson 2021: 509.) Given that the relation that matters is temporally ordered—that is, like Relation *R* rather than Relation *R*′—each person-slice has an incitement (given a prudential motivation) to get another scan done and create even more replicas (which fuels the explosion of replicas further).

[57] Yet it may be harder to merge than to split. See Hanson (2016: 51).

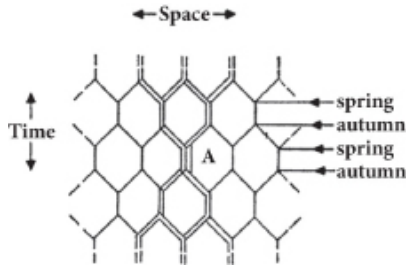


Figure 4.1 Regular intervals of merging and splitting.

remembers) earlier person-slices from multiple branches.[58] Then each simulation that will merge is *R*-related to all descendants of the merged simulation. This solution is structurally the same as Parfit's example of beings who merge and divide every autumn and spring (Figure 4.1).[59] With these regular intervals of merging and splitting, everyone's prudential interests would overlap to a very large extent with those of everyone else.

So far in our discussion of uploading, we have assumed that psychology is what matters in survival. If (i) psychological continuity is what matters, (ii) uploading is feasible, (iii) uploading preserves psychological continuity, and (iv) simulations of brains would be conscious, it follows that one would have what matters in survival to a simulation from one's uploaded brain. Or, at least, it follows if this continuity need not have its normal cause: being caused by the continued existence of one's brain.[60]

But there is at least one rival to the psychological view that may also allow that uploading provides what matters. On the phenomenal view of what matters in survival, someone has what matters in relation to a future person-slice if and only if they are phenomenally continuous with that future person-slice.[61] Phenomenal continuity is, basically, the relation of partaking of the same stream of consciousness. In the same way as psychological continuity is the holding of overlapping sequences of psychological connectedness, phenomenal continuity is the holding of overlapping sequences of phenomenal connectedness. Phenomenal connectedness, in turn, is the relation of experienced togetherness—that is, the relation of one experience being experienced together with another experience in the same conscious state. This relation can hold between experiences at a time, such as your current visual experience and your current auditory experience. But this relation of experienced togetherness can also hold over time. For example, when you are listening to music and you hear one note transition into the next, or more generally, each one of your experiences flows into the next.[62] On this phenomenal view, a simulation provides what matters in survival for a person *S* if and only if a stream of consciousness which *S* currently partakes of will include the experiences of the simulation.

[58] Bostrom (2014: 61) observes that digital minds might want to share memories to increase their knowledge faster. They may be able to save computational resources by storing just one instance of the memories, even though they can all access those memories. This raises the question of where a mind ends. See Clark and Chalmers (1998).

[59] Parfit (1971: 22; 1984: 303).

[60] Parfit (1984: 283–7) defends this kind of view where there are no restrictions on how the psychological continuity is caused.

[61] Gustafsson (2021: 513 fn. 28).

[62] Dainton and Bayne (2005: 553–4). Phenomenal continuity can also hold between experiences that are separated by a period of unconsciousness, such as dreamless sleep. See Gustafsson (2011: 291–4).

If simulations can be conscious and have experiences at all, could we get a human person's stream of consciousness to transfer to a simulation? Barry Dainton suggests that we can. By gradually replacing more and more of the person's brain with functionally equivalent digital silicon-based parts, he suggests that their stream of consciousness would continue intact.[63] Then, if done gradually, uploading may still provide what matters in survival even on the phenomenal view.

4 Biological uploading

One worry about uploading is that, as mentioned earlier, computer simulations might not, for all we know, be conscious. Yet, if we have a detailed scan of someone's brain, we might be able to create a new human complete with their psychology through what we will call *biological uploading*. The standard implementation of this idea is the teletransporter: A person steps into a machine on Earth which scans their body and then eliminates it and sends the scanned information to Mars, where a biological copy of the person is generated from the scan.[64]

If the creation of biological replicas with someone's psychology is possible, it may lead to Prudential Longtermism in much the same way as uploading. Or, at least, it may do so on the psychological view of what matters in survival. By contrast, on the phenomenal view, it seems less likely that the scanned person would have what matters in relation to their replicas, because it's implausible that the replicas' experiences would be part of the same stream of consciousness as the scanned person's experiences.

But we may be able to use a similar gradual approach to get it to work. It seems that, if a person *S*'s brain is split in two, *S* would have phenomenal continuity to both halves. These brain halves can then be placed in two separate bodies and complemented with a replica of the other half. The result should be two people who each have a complete brain and who are both phenomenally continuous with *S*. Then we repeat this, if necessary, to generate the desired number of replicas.

(Actually, if we don't care about keeping *S's* psychology intact, this last procedure does not, fundamentally, need uploading. If we don't care about psychology, the brain halves needn't be combined with replicas of the other half—any compatible brain half will do.)

5 Prudential and empirical uncertainty

So far, we have seen that uploading and biological uploading can lead to Strong Prudential Longtermism. This requires that either psychological or phenomenal continuity is what matters in survival and, moreover, that we aggregate well-being in the way prescribed by the Prudential Total View (or some similar additive principle). These assumptions are plausible. The prudential analogue of the Mere-Addition Paradox is compelling and suggests the Prudential Total View or something similarly additive. That the relation that matters

[63] Or rather, on Dainton's (2012: 55) view, the person's *capacity* for a continuous stream of consciousness would continue intact. See also Chalmers (2010: 52–5) for a discussion of gradual uploading.

[64] Parfit (1984: 199).

in survival is some mental relation (psychological or phenomenal) is also compelling. The reason why our brains and bodies seem so important in survival is that they are needed (so far) for mental continuity—but they are not what fundamentally matters. Even so, few of us are *certain* that all of these assumptions are true. To handle uncertainty regarding these normative questions about what matters, maximizing expected prudential value is analogous to maximizing expected moral value in decisions under moral uncertainty.[65]

But we also have descriptive uncertainty. The technologies we need in order to implement these approaches have yet to be invented, and it's unclear when they will, if ever. But the two technologies that did not (by themselves) lead to Strong Prudential Longtermism (that is, anti-ageing and cryonics) might still buy us time for uploading or biological uploading to become feasible. Especially anti-ageing seems promising, as anti-ageing research has made some significant advances in recent years.[66] If anti-ageing works, it might raise our life expectancy by several hundred years. This should, then, give us time to perfect either uploading or biological uploading.[67]

If either uploading or biological uploading becomes technologically feasible and our assumptions about what matters are correct, then uploading could create an enormous amount of prudential value through longevity, fission, and increasing quality of life. So, even if there were only a small chance that these technologies will work during the lifetimes of some currently existing people and we are uncertain whether our assumptions about

[65] See Lockhart (2000: 82). Maximizing expected prudential value is open to the same worry about intertheoretic comparisons of value. See Ross (2006: 761–5) and Gustafsson and Torpman (2014: 160–5). Some alternative approaches avoid intertheoretic comparisons of value, for example: My Favourite Theory (Gracely 1996: 331; Gustafsson and Torpman 2014: 167–70), My Favourite Option (Lockhart 1992: 35–6), the Borda Rule (MacAskill 2016: 989; MacAskill, Bykvist, and Ord 2020: 73). But these alternative approaches will, in some cases of predicted future moral progress, lower the expected moral value conditional on every moral theory in which we have any credence. See Gustafsson (2022b: 452–66). So it seems that we have to, as well as we can, rely on intertheoretic comparisons of value. Still, just like average and total utilitarianism lack a common unit, the average and total views for aggregation in fission cases and within life-paths also lack a common unit. See Broome (2012: 185).

[66] See de Grey and Rae (2007: 49–308) and Partridge et al. (2020).

[67] One worry, however, is a prudential analogue of *the Doomsday Argument*—a notorious argument that, taking ourselves to be a random sample from all people in history, we should, given our relatively early position, lower our credence in that humanity will colonize the stars and create an enormous number of future people. (See Leslie 1989: 10; 1996: 187–236; Bostrom 2002: 89–108.) What we may call *the Prudential Doomsday Argument* is an analogous argument that you won't live for an extremely long time. See Korb and Oliver (1998: 405 fn. 2), and, for a similar one-person argument against the likelihood of an eternal afterlife, see Leslie (2008: 520–4) and Page (2010: 397–401). If your life will be extremely long (or split into an enormous amount of uploads), then most of your observer-moments will be observer-moments where you are much older than you are now, or they will be simulated observer-moments. We apply

> *The Strong Self-Sampling Assumption*: One should reason as if one's present observer-moment was a random sample from the set of all observer-moments in its reference class.

(Bostrom 2002: 126.) We take the reference class for your current observer-moment to include all your observer-moments. So, if you regard your current observer-moment as a random sample from all of your observer-moments, it would be surprising if you got an observer-moment where you are still this young and not a simulation. So, the argument goes, you should consider it unlikely that you will live for an extremely long time or split into an enormous amount of simulations (assuming that you can tell whether you are simulated). Or, at least, you should regard this possibility as less likely than you did before you considered the Prudential Doomsday Argument. Bostrom (2002: 111–5) objects that the relevant reference class should include not only your observer-moments but also all other observer-moments. If so, he argues, the Prudential Doomsday Argument falls apart because, once we take into account that long lives include more observer-moments, we neutralize the adjustment for finding that your current observer-moment is early. But his solution assumes that we already know the average lifespan of the people in our reference class. In our discussion of the feasibility of extreme life extension, this isn't something we know in advance. Moreover, a standard defence against the Doomsday Argument is to adopt

> *The Self-Indication Assumption*: Given the fact that you exist, you should (other things equal) favour hypotheses according to which many observers exist over hypotheses on which few observers exist.

what matters are correct, we should still get that Strong Prudential Longtermism holds for some currently existing people in terms of their overall expectation of prudential value.[68] This is fully consistent with these technologies being unlikely to work.

6 Longtermism based on Prudential Longtermism

Given Prudential Longtermism, a large number of theories that otherwise wouldn't lead to (impersonal) Longtermism may turn out to do so.[69] Person-affecting views on which we should minimize the strongest complaint would lead to Longtermism.[70] This is so, since the strongest complaints will come from people for whom Prudential Longtermism is true. Likewise, common-sense morality, on which one should prioritize one's family and friends, would lead to Longtermism if Prudential Longtermism holds for a sufficient number of one's family and friends. Self-interest theories would lead to Longtermism if Prudential Longtermism holds for the agent. Finally, person-affecting utilitarianism would lead to Longtermism if Prudential Longtermism holds for a sufficient number of current people.[71]

The practical implications of Longtermism based on Prudential Longtermism would differ in some respects from those of Longtermism based on total utilitarianism. In addition to prioritizing the reduction of existential risk to safeguard humanity as a whole, Longtermism based on Prudential Longtermism would also prioritize speeding up technological progress in the areas that might help life extension.[72] It would prioritize funding life extension, so that, in the long run, some of us may still be alive.[73] (Compared to regular Longtermism, Prudential Longtermism would recommend being more willing to take greater existential risks with AI since fast AI progress plausibly increases the chance of developing life extension in time.) The badness of death plausibly consists, largely, in how much better a person's life would have been in expectation if they had lived on.[74] Consequently, Prudential Longtermism makes avoiding an early death all the more pressing.[75]

(Bostrom 2002: 66.) This principle neutralizes the Doomsday Argument. (See Bostrom 2002: 122–3.) The trouble is that this kind of move does not seem very plausible against the Prudential Doomsday Argument. The analogue of the Self-Indication Assumption for observer-moments would be

> *The Strong Self-Indication Assumption*: Given the fact that you have a current observer-moment, you should (other things being equal) favour hypotheses according to which many observer-moments exist over hypotheses on which few observer-moments exist.

But this principle no more favours hypotheses with lots of long lives than hypotheses with the same number of observer-moments but with only short lives. For an objection based on the idea that the first moment one considers the Prudential Doomsday Argument is likely to come early even in a long life, see van Inwagen (2016: 217–18).

[68] Maximizing expected prudential value may seem to lead to a kind of fanaticism in these kinds of cases where the overall calculation is dominated by a very unlikely but enormously valuable outcome. See Smith (2014), Monton (2019), and Kosonen (2022: 137–239). But deviations from expected utility theory are vulnerable to money pumps. See Gustafsson (2022a) and Kosonen (2022: 196–239).

[69] Prudential longtermism is fairly implausible for non-human animals. So the case for Longtermism based on Prudential Longtermism may be weaker on views where non-human animals typically dominate the overall calculation of value. But, once we take Prudential Longtermism into account, it may be that non-human animals no longer dominate.

[70] Parfit's (n.d.: ch. 6) principle '*Minimax Loss*: The best outcome is the one in which the greatest loser loses least.'

[71] Bostrom (2003: 311–2).

[72] Bostrom (2003: 313–4).

[73] Compare Keynes's (1923: 80) more pessimistic assessment.

[74] Broome (1993: 83).

[75] We wish thank Jacob Barrett, Tim Campbell, Tomi Francis, Hilary Greaves, Todd Karhu, Kevin Kuruc, Andreas Mogensen, Christian Tarsney, Teru Thomas, and David Thorstad for valuable comments.

References

Aaronson, S. (2016), 'The Ghost in the Quantum Turing Machine', in S. B. Cooper and A. Hodges (eds.), *The Once and Future Turing: Computing the World* (Cambridge University Press), 193–296.

Ahmed, A. (2014), *Evidence, Decision and Causality* (Cambridge University Press).

Ahmed, A. (2020), 'Rationality and Future Discounting', in *Topoi* 39/2: 245–56.

Arias, E. and Xu, J. (2022), 'United States Life Tables, 2019', in *National Vital Statistics Reports* 70/19: 1–58.

Arrhenius, G. (2000), *Future Generations: A Challenge for Moral Theory*, PhD thesis, Uppsala University.

Arrhenius, G. and Rabinowicz, W. (2015), 'The Value of Existence', in I. Hirose and J. Olson (eds.), *The Oxford Handbook of Value Theory* (Oxford University Press), 424–43.

Benford, G. et al. (2005), 'Scientists' Open Letter on Cryonics', https://www.biostasis.com/scientists-open-letter-on-cryonics/ (access date September 1, 2020).

Bostrom, N. (2002), *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge).

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', in *Utilitas* 15/3: 308–14.

Bostrom, N. (2005), 'The Fable of the Dragon-Tyrant', in *Journal of Medical Ethics* 31/5: 273–7.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

Bostrom, N. and Ord, T. (2006), 'The Reversal Test: Eliminating Status Quo Bias in Applied Ethics' in *Ethics* 116/4: 656–79.

Bostrom, N. and Roache, R. (2007), 'Human Enhancement: Ethical Issues in Human Enhancement', in J. Ryberg, T. S. Petersen, and C. Wolf (eds.), *New Waves in Applied Ethics* (Palgrave Macmillan), 120–52.

Brink, D. O. (1992), 'Sidgwick and the Rationale for Ratinal Egoism', in B. Schultz (ed.), *Essays on Henry Sidgwick* (Cambridge University Press), 199–240.

Broome, J. (1993), 'Goodness Is Reducible to Betterness: The Evil of Death Is the Value of Life', in P. Koslowski and Y. Shionoya (eds.), *The Good and the Economical: Ethical Choices in Economics and Management* (Springer), 70–84.

Broome, J. (2004), *Weighing Lives* (Oxford University Press).

Broome, J. (2012), *Climate Matters: Ethics in a Warming World* (Norton).

Chalmers, D. J. (2010), 'The Singularity: A Philosophical Analysis', in *Journal of Consciousness Studies* 17/9–10: 7–65.

Clark, A. and Chalmers, D. (1998), 'The Extended Mind', in *Analysis* 58/1: 7–19.

Crimmins, E. M. (2015), 'Lifespan and Healthspan: Past, Present, and Promise', in *The Gerontologist* 55/6: 901–11.

Crisp, R. (1997), *Mill on Utilitarianism* (Routledge).

Dainton, B. (2012), 'On Singularities and Simulations', in *Journal of Consciousness Studies* 19/1–2: 42–85.

Dainton, B. and Bayne, T. (2005), 'Consciousness as a Guide to Personal Persistence', in *Australasian Journal of Philosophy* 83/4: 549–71.

de Grey, A. and Rae, M. (2007), *Ending Aging: The Rejuvenation Breakthroughs That Could Reverse Human Again in Our Lifetime* (St. Martin's Press).

Doyle, D. J. (2018), *What Does It Mean to Be Human? Life, Death, Personhood and the Transhumanist Movement* (Springer).

Dyson, F. J. (1979), 'Time Without End: Physics and Biology in an Open Universe', in *Reviews of Modern Physics* 51/3: 447–60.

Gibbard, A. and Harper, W. L. (1978), 'Counterfactuals and Two Kinds of Expected Utility', in C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory*, vol. I, (Reidel), 125–62.

Gracely, E. J. (1996), 'On the Noncomparability of Judgments Made by Different Ethical Theories', in *Metaphilosophy* 27/3: 327–32.

Greaves, H. (2019), 'Review of Samuel Scheffler, *Why Worry about Future Generations?*', in *Ethics* 130/1: 136–41.

Greaves, H. and MacAskill, W. (this volume), 'The Case for Strong Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).

Gustafsson, J. E. (2011), 'Phenomenal Continuity and the Bridge Problem', in *Philosophia* 39/2: 289–96.

Gustafsson, J. E. (2018), 'The Unimportance of Being Any Future Person', in *Philosophical Studies* 175/3: 745–50.

Gustafsson, J. E. (2019), 'Non-Branching Personal Persistence', in *Philosophical Studies* 176/9: 2307–29.

Gustafsson, J. E. (2021), 'Is Psychology What Matters in Survival?', in *Australasian Journal of Philosophy* 99/3: 504–16.

Gustafsson, J. E. (2022a), *Money-Pump Arguments* (Cambridge University Press).

Gustafsson, J. E. (2022b), 'Second Thoughts about My Favourite Theory', in *Pacific Philosophical Quarterly* 103/3: 448–70.
Gustafsson, J. E. and Kosonen, P. (2024), 'Do Lefty and Right Matter More Than Lefty Alone?', in *Erkenntnis* 89/5: 1921–1926.
Gustafsson, J. E. and Torpman, O. (2014), 'In Defence of My Favourite Theory', in *Pacific Philosophical Quarterly* 95/2: 159–74.
Hanson, R. (2016), *The Age of Em: Work, Love, and Life When Robots Rule the Earth* (Oxford University Press).
Holtug, N. (2001), 'The Repugnant Conclusion about Self-Interest', in *Danish Yearbook of Philosophy* 36/1: 49–68.
Holtug, N. (2010), *Persons, Interests, and Justice* (Oxford University Press).
Jeffrey, R. C. (1965), *The Logic of Decision* (McGraw-Hill).
Jeffrey, R. C. (1983), *The Logic of Decision*, 2nd edition (University of Chicago Press).
Keynes, J. M. (1923), *A Tract on Monetary Reform* (Macmillan).
Korb, K. B. and Oliver, J. J. (1998), 'A Refutation of the Doomsday Argument', in *Mind* 107/426: 403–10.
Kosonen, P. (2022), *Tiny Probabilities of Vast Value*, PhD thesis, University of Oxford.
Leslie, J. (1989), 'Risking the World's End', in *Bulletin of the Canadian Nuclear Society* 10/3: 10–15.
Leslie, J. (1996), *The End of the World: The Science and Ethics of Human Extinction* (Routledge).
Leslie, J. (2008), 'Infinitely Long Afterlives and the Doomsday Argument', in *Philosophy* 83/326: 519–24.
Lewis, D. (1976), 'Survival and Identity', in A. O. Rorty (ed.), *The Identities of Persons* (University of California Press), 17–40.
Lewis, D. (1986), *On the Plurality of Worlds* (Blackwell).
Lockhart, T. (1992), 'Professions, Confidentiality, and Moral Uncertainty', in *Professional Ethics* 1/3–4: 33–52.
Lockhart, T. (2000), *Moral Uncertainty and Its Consequences* (Oxford University Press).
MacAskill, W. (2016), 'Normative Uncertainty as a Voting Problem', in *Mind* 125/500: 967–1004.
MacAskill, W. (2019), 'Longtermism', Effective Altruism Blog, https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism (access date August 1, 2020).
MacAskill, W, Bykvist, K., and Ord, T. (2020), *Moral Uncertainty* (Oxford University Press).
McMahan, J. (1981), 'Problems of Population Theory', in *Ethics* 92/1: 96–127.
McMahan, J. (2002), *The Ethics of Killing: Problems at the Margin of Life* (Oxford University Press).
McTaggart, J. M. E. (1927), *The Nature of Existence Volume II* (Cambridge University Press).
Merkle, R. C. (1992), 'The Technical Feasibility of Cryonics', in *Medical Hypotheses* 39/1: 6–16.
Merkle, R. C. (1994), 'The Molecular Repair of the Brain, Part I', in *Cryonics* 15/1: 16–31.
Miller, K. (2004), 'Cryonics Redux: Is Vitrification a Viable Alternative to Immortality as a Popsicle?', in *Skeptic* 11/1: 24–5.
Minerva, F. (2018), *The Ethics of Cryonics: Is It Immoral to Be Immortal?* (Palgrave Macmillan).
Mogensen, A. (this volume), 'Would a World Without Us Be Worse?', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Monton, B. (2019), 'How to Avoid Maximizing Expected Utility', in *Philosophers' Imprint* 19/18: 1–24.
Nagel, T. (1986), *The View from Nowhere* (Oxford University Press).
Narveson, J. (1973), 'Moral Problems of Population', in *Monist* 57/1: 62–86.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Page, D. N. (2010), 'Scientific and Philosophical Challenges to Theism' in M. Y. Stewart (ed.), *Science and Religion in Dialogue* (Blackwell), 396–410.
Parfit, D. (1971), 'Personal Identity', in *The Philosophical Review* 80/1: 3–27.
Parfit, D. (1984), *Reasons and Persons* (Clarendon Press).
Parfit, D. (1986), 'Overpopulation and the Quality of Life', in P. Singer (ed.), *Applied Ethics* (Oxford University Press), 145–64.
Parfit, D. (1993), 'The Indeterminacy of Identity: A Reply to Brueckner', in *Philosophical Studies* 70/1: 23–33.
Parfit, D. (1995), 'The Unimportance of Identity', in H. Harris (ed.), *Identity: Essays Based on Herbert Spencer Lectures Given in the University of Oxford* (Clarendon Press), 13–45.
Parfit, D. (2017), *On What Matters: Volume Three* (Oxford University Press).
Parfit, D. (n.d.), *On Giving Priority to the Worse Off* (unpublished manuscript).
Partridge, L., Fuentealba, M., and Kennedy, B. K. (2020), 'The Quest to Slow Ageing Through Drug Discovery', in *Nature Reviews Drug Discovery* 19/8: 513–32.
Perry, J. (1972), 'Can the Self Divide?', in *The Journal of Philosophy* 69/16: 463–88.
Ramakrishnan, V. (2024), *Why We Die: The New Science of Aging and the Quest for Immortality* (William Morrow).

Ramsey, F. P. (1928), 'A Mathematical Theory of Saving', in *The Economic Journal* 38/152: 534–59.
Rawls, J. (1971), *A Theory of Justice* (Harvard University Press).
Rawls, J. (1999), *A Theory of Justice*, revised edition (Harvard University Press).
Ross, J. (2006), 'Rejecting Ethical Deflationism', in *Ethics* 116/4: 742–68.
Ross, J. (2014), 'Divided We Fall', in *Philosophical Perspectives* 28/1: 222–62.
Sagan, C. (1994), *Pale Blue Dot: A Vision of the Human Future in Space* (Random House).
Sandberg, A. and Bostrom, N. (2008), *Whole Brain Emulation: A Roadmap*, Technical Report 2008-3 (Future of Humanity Institute).
Scheffler, S. (2013), *Death and the Afterlife* (Oxford University Press).
Scheffler, S. (2018), *Why Worry about Future Generations?* (Oxford University Press).
Shaw, D. (2009), 'Cryoethics: Seeking Life After Death', in *Bioethics* 23/9: 515–21.
Shoemaker, S. (1970), 'Persons and Their Pasts', in *American Philosophical Quarterly* 7/4: 269–85.
Sidgwick, H. (1907), *The Methods of Ethics*, 7th edition (Macmillan).
Smith, N. J. J. (2014), 'Is Evaluative Compositionality a Requirement of Rationality?', in *Mind* 123/490: 457–502.
Steele, K. (this volume), 'Longtermism and Neutrality about More Lives', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Tappenden, P. (2011), 'Expectancy and Rational Action Prior to Personal Fission', in *Philosophical Studies* 153/2: 299–306.
Temkin, L. S. (1987), 'Intransitivity and the Mere Addition Paradox', in *Philosophy & Public Affairs* 16/2: 138–87.
Temkin, L. S. (2008), 'Is Living Longer Living Better?', in *Journal of Applied Philosophy* 25/3: 193–210.
Temkin, L. S. (2012), *Rethinking the Good* (Oxford University Press).
Thomas, T. (2023), 'The Asymmetry, Uncertainty, and the Long Term', in *Philosophy and Phenomenological Research* 7/2: 470–500.
Unger, P. (1990), *Identity, Consciousness and Value* (Oxford University Press).
van Inwagen, P. (2016), 'The Rev'd Mr Bayes and the Life Everlasting', in M. Bergmann and J. E. Brower (eds.), *Reason and Faith: Themes from Richard Swinburne* (Oxford University Press), 196–219.
Vita-More, N. and Barranco, D. (2015), 'Persistence of Long-Term Memory in Vitrified and Revived *Caenorhabditis Elegans*', in *Rejuvenation Research* 18/5: 458–63.
Williams, B. (1973), 'The Makropulos Case: Reflections on the Tedium of Immortality', in *Problems of the Self: Philosophical Papers 1956–1972* (Cambridge University Press), 82–100.

5

Would a World Without Us Be Worse?

Clues from Population Axiology

Andreas L. Mogensen

1 Introduction

Many think that human extinction would be a catastrophic tragedy, and that reducing extinction risk should be among our highest priorities (Parfit 2011; Ord 2020). Not everyone is so sure. Some take the view that human lives are not worth starting because of the suffering they contain (Benatar 2006). Scepticism about the value of continued human survival is especially likely to arise among those concerned about our effects on non-human animals and the natural world (Taylor 1981; May 2018).

This paper engages with the question of whether and to what extent the continued existence of humanity is morally desirable. In addressing this question, we inevitably have to engage with questions of *population axiology*—questions about how to value outcomes that differ in the size and/or composition of the population (Greaves 2017). This is notoriously hard (Parfit 1984; Ng 1989; Carlson 1998; Arrhenius 2000; Kitcher 2000). As a result, it's difficult to be confident in any one theory, and natural to look across a range of theories. That's the goal of this chapter.

Section 2 briefly considers a view with an especially infamous relationship to the desirability of extinction: *negative utilitarianism*. Section 3 considers a different view that bears important similarities to negative utilitarianism: *the procreation asymmetry*. I explain why the procreation asymmetry does not support the desirability of extinction, contrary to appearances. Section 4 is on *total utilitarianism*, with particular focus on versions of total utilitarianism that avoid the Repugnant Conclusion. Section 5 is on *critical-level utilitarianism*. Section 6 is on *average utilitarianism* and *variable value theories*. The sections from here on discuss axiologies that disagree with utilitarian theories even in fixed population cases, but with a specific focus on how these views might affect the evaluation of population changes of the kind relevant to assessing continued human survival. Section 7 is on *prioritarianism* and *egalitarianism*, where I primarily consider an argument that the value of equality counts for the desirability of human extinction. Moving beyond welfarist axiologies, section 8 is on *perfectionism*, where I consider the possibility that the prospect of cultural decline functions as a reason to hasten extinction given a perfectionist axiology. Lastly, in section 9, I consider *conservatism about value* and explain why conservatism provides less support for the desirability of human survival than has been claimed.

Andreas L. Mogensen, *Would a World Without Us Be Worse?* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Andreas L. Mogensen 2025. DOI: 10.1093/9780191979972.003.0005

2 Negative utilitarianism

No theory of value bears a more infamous relationship to the desirability of human extinction than *negative utilitarianism* (NU). Popper (1945 [2011]: 602) claims that 'pain cannot be outweighed by pleasure' and that our goal should be 'the least amount of avoidable suffering for all'. These remarks may be interpreted as follows.[1] We assume that welfare has both a positive aspect and a negative aspect. Call the former *well-being* and the latter *ill-being*. NU says that one outcome, *x*, is worse than another, *y*, if the sum of ill-being is greater in *x*.[2]

The best-known objection to NU is that it entails the moral desirability of human extinction. Smart (1958) asks us to imagine a weapon capable of instantly and painlessly killing every human being. He argues that since everyone would experience some suffering if they live on, NU counts the painless death of every human being now living as better than their continued existence.

However, this argument is not strictly valid. It neglects non-human welfare subjects (Knutsson 2019). Human extinction would minimize the sum of ill-being experienced by human beings, but cannot be assumed to minimize the sum of ill-being considered in full generality without further argument. What Smart should have said, arguably, is that NU entails the desirability of the extinction of all welfare subjects.[3] It does not obviously entail the desirability of human extinction except as part of such an event.

3 The procreation asymmetry

A well-known view in population ethics with similarities to NU is *the procreation asymmetry* (PA), the view that whereas adding lives that are not worth living to the population makes the outcome worse, *ceteris paribus*, adding lives that are worth living to the population does not make the outcome better (or worse), *ceteris paribus* (McMahan 1981; Roberts 2011). Just as NU attaches no value to well-being, but only disvalue to ill-being, PA attaches no value to additional lives in which well-being predominates, but only disvalue to additional lives in which ill-being predominates.

Unsurprisingly, therefore, PA has also been claimed to favour extinction (Sikora 1978; Rachels 1998; Holtug 2004). Holtug (2004) asks us to imagine that people can either choose to carry on the human race or let it go extinct. PA supposedly entails that the latter would be better, 'because, among the billions of people they could cause to exist, there would surely be a few . . . who would be miserable; and while their misery would count against their being created, the happiness of the rest would count for nothing' (Holtug 2004: 139).

[1] Note that Popper (1945 [2011]: 501) would subsequently forswear this interpretation. Thanks to Jacob Barrett for noting this.

[2] Consistent with Popper's remarks, we may also count *x* as worse than *y* if the sum of ill-being is equal in *x* and *y*, but the sum of well-being is greater in *y*. Thanks to Hilary Greaves for this observation.

[3] In fairness, he comes very close. While Smart (1958) does pose his objection to NU in terms of 'a weapon capable of . . . destroying the human race' (542), he also describes whoever wields his imagined weapon as a 'world-exploder' (543) and suggests that, given NU, such a person is 'the saviour of mankind, and for that matter of the animals too' (543).

This argument may also be criticized for neglecting non-human welfare subjects.[4] Even apart from this, it is unsound. The desirability of extinction follows from PA in the way Holtug suggests only if we assume that a bad thing plus a neutral thing adds up to a bad thing. While intuitive, it is well known that those who defend PA have powerful independent reasons to reject this assumption and posit that the neutrality of additional good lives is 'greedy', being 'able to swallow up bad things and neutralize them' (Broome 2005: 409).

Here is why. PA entails *the principle of neutrality* (PN), which says that one population that differs from another only in that it involves any number of additional lives that are worth living is not better or worse than the status quo population. This could be taken to mean that the smaller population is equal in value to the larger. This leads to very bad results, as demonstrated by Broome (2005). A more plausible interpretation of PN is as the view that one population that differs from another only in that it involves any number of additional good lives is neither better than, worse than, nor exactly as good as the smaller population.

However, Broome shows that when PN is so interpreted, we are forced to conclude that one outcome that differs from another in two respects—one bad, the other neutral—need not be worse overall. Consider the following possible outcomes, where a numerical value in a cell denotes a person's lifetime welfare in the corresponding population, whereas 'Ω' denotes their non-existence:

	Afryea	**Beom-seok**
i	5	Ω
j	5	1
k	4	4

Plausibly, *k* is better than *j*. This follows from *non-antiegalitarianism* (NAE), according to which, if the same people exist in *x* and *y* and *y* is perfectly equal with higher total and average welfare than *x*, then *y* is better than *x*, *ceteris paribus* (Ng 1989). PN entails that *j* is not worse than *i*. It follows that *k* is not worse than *i*. This is surprising: *k* differs from *i* in just two respects, one of which is bad (the loss to Afryea of one unit of welfare) and one of which is assumed to be neutral (the addition of Beom-seok).

A priori, one might have thought that a bad thing plus a neutral thing must add up to a bad thing. Frick (2017) calls this the *decomposition principle* (DP). Anyone who endorses PA, and *ipso facto* PN, has good reason to reject DP. Broome thinks DP is compelling and rejects PN.[5] Many of his respondents believe that DP may reasonably be jettisoned instead (Rabinowicz 2009; Frick 2017; Nebel 2019). For present purposes, we can argue as follows.

[4] Strictly speaking, Holtug (2004: 139) asks us to imagine the choice described in the previous paragraph as one confronted 'sometime in the future, [by] the last few inhabitants of earth'. Thus, it may be that in this scenario we are to imagine that there exist only those human beings who must choose whether to carry on their species.

[5] Broome also argues against PN on the grounds that if the value of additional lives worth living is 'greedy', then we are not able to capture the intuitions taken to support PN, since we are not typically able to ignore potential changes in population size when deciding how to act, given the ability of changes in the number of lives worth living to cancel out other morally significant changes we might bring about.

The argument that PA entails the desirability of extinction fails, since it implicitly relies on DP, whereas those who endorse PA have strong independent reason to reject DP.[6]

4 Total utilitarianism

The next three sections consider a range of well-known, non-negative utilitarian population axiologies that all reject PN, beginning with *total utilitarianism* (TU). For two outcomes, x and y, with $N(x)$ and $N(y)$ individuals, respectively, if $u_i(x)$ denotes the lifetime welfare of individual i in x ($i = 1,\ldots,N(x)$) and $u_i(y)$ denotes the lifetime welfare of individual i in y ($i = 1,\ldots,N(y)$), then, on TU, x is at least as good as y if and only if

$$\sum_{i=1}^{N(x)} u_i(x) \geq \sum_{i=1}^{N(y)} u_i(y)$$

In other words, we compare the sum of each individual's welfare for the two outcomes and regard the first as at least as good as the second just in case this sum is at least as high in the first outcome as in the second.

TU has many desirable properties. For example, unlike average utilitarianism, it is separable, and so allows us to ignore the kind of exobiological issues we'll eventually have to face in section 6. However, much of the literature on population ethics focuses on counter- intuitive implications of this view, especially the claim that it entails the *Repugnant Conclusion* (RC) (Parfit 1984): for any population, A, of lives with very high welfare levels, there is a population, Z, of lives that are barely worth living that is better than A. Are trade-offs between lives of the kind under comparison in the statement of RC relevant in thinking about the value of continued human survival?

Plausibly, they are. As noted, scepticism about the value of continued human survival is especially likely to arise among those concerned about the negative impact of human beings on non-human animals and the natural world. Terrestrial vertebrate species have seen a mean decline of 28% in numbers over the past 40 years, and marine vertebrates have declined in abundance by 22% on average (Dirzo et al. 2014; McCauley et al. 2015). Alongside our treatment of factory-farmed animals, these are among the kind of considerations highlighted by May (2018) in making the case that we should take seriously the desirability of human extinction.

Is there a case to be made for the desirability of human extinction in light of the fact that human beings so thoroughly decimate wild animal populations? Such a case will be hard to make if the members of the wild animal species that suffer population declines due to human activity do not have lives worth living. (More on this later.) Assume, therefore, that their lives *are* worth living. Someone might then claim that the choice between a large human population and the much larger population of wild animals that may be projected to exist in our absence involves one important aspect of the choice between an A population and a Z population. For reasons to be explained immediately below, this comparison may

[6] For more on the implications of PN for longtermism, see Steele (this volume).

be thought to involve, at least in part, a choice between a large population of individuals with high lifetime welfare levels and a much larger population of individuals whose lives are very near the neutral level.

Plausibly, many human lives are very good. Globally, around 32% of human beings report that they are very happy, rising as high as 67% in Mexico (Inglehart et al. 2014).[7] Human lives may be thought of as ennobled by higher goods inaccessible to most or all non-human animals, such as autonomy, artistic creativity, and scientific understanding. The lives of non-human animals are typically much, much shorter than those of human beings and lived in very hard conditions, characterized by scarcity of food and water and vulnerability to predation as normal parts of life (Tomasik 2015). Minute by minute, 'thousands of animals are being eaten alive, many others are running for their lives, whimpering with fear, others are slowly being devoured from within by rasping parasites, thousands of all kinds are dying of starvation, thirst, and disease' (Dawkins 1995: 132). An individual whose lifespan, cognitive abilities, and affect balance are like those of a typical non-human animal living in the wild might be thought to have a life scarcely better, if at all, than a life like that originally imagined by Parfit (1986: 148) for the Z population in RC, a population in which 'people . . . never suffer; but all they have is muzak and potatoes'.[8]

The overall picture sketched here is obviously contestable. The claims made about the welfare of wild animals may seem overly pessimistic. I also haven't taken account of the possibility that shrinking wild animal populations may reduce the expected human population over all time due to the threat of ecosystem collapse (Dirzo, Ceballos, and Ehrlich 2022). It might also be claimed that the total population of wild animals over all time is greater in expectation given continued human survival because only the technological capabilities of our descendants can safeguard the biosphere against total destruction due to the brightening of the Sun in roughly a billion years' time (Ord 2020: 218–223). Nonetheless, it seems far from absurd that questions about the value of continued human survival that arise in response to the current defaunation crisis involve comparisons of the kind that are at play in RC.

On the other hand, Thomas (2018) and Nebel (2022) have recently brought renewed attention to the possibility that TU, suitably interpreted, need not entail RC (compare Portmore 1999; Kitcher 2000; Carlson 2007). TU ranks outcomes by comparing $\sum_{i=1}^{N(x)} u_i(x)$ and $\sum_{i=1}^{N(y)} u_i(y)$. We have assumed implicitly that $u_i(x)$ and $u_i(y)$ are scalar quantities. Suppose instead that they are vectors of real numbers of the form (a,b) We let a be a measure of the dimension of an individual's welfare corresponding to something like Mill's 'higher pleasures' and b be a measure of the dimension of that individual's welfare corresponding to the 'lower pleasures'. Vectors can be added in the standard piecewise fashion: $(a,b)+(c,d)=(a+c,\, b+d)$. They can also be ordered by the standard lexicographic ordering: $(a,b) \geq (c,d)$ just in case $a > c$ or $a = c$ and $b \geq d$.[9] Thus, vector-valued representations of individual welfare levels are compatible with TU.

Call a view of this kind VTU. Assume that each dimension of welfare has a zero level. Say that an individual's life is neutral if $a = 0$ and $b = 0$, barely worth living if $a = 0$ and $b > 0$,

[7] Overall, 83.6% of respondents in Inglehart et al. (2014) report that they are either very happy or quite happy.

[8] Notably, Parfit (2016: 118) later asks us to imagine the Z population in RC as composed of animals that have only 'enough slight pleasures like those of cows munching grass or lizards basking in the sun'.

[9] Nebel revises the standard lexicographic ordering in order to avoid certain objections. On the revised ordering, there exist $\Delta, \delta > 0$, such that for quantities of welfare (a,b), (c,d), we have $(a,b) \geq (c,d)$ just in case (i) $a - c > \Delta$, (ii) $a \geq c$ and $b \geq d$, or (iii) $a \geq c$ and $(a-c)/(d-b) > \Delta/\delta$.

and good if $a > 0$. So understood, VTU avoids RC, in that a population with good lives corresponding to the welfare level represented as (100, 0) is better than any population with lives barely worth living corresponding to a welfare level represented as $(0, \epsilon)$ for any $\epsilon > 0$.

VTU also turns out to violate a principle that bears in a different way on the potential desirability of human extinction, given the harms we inflict on non-human animals. This principle says, roughly, that the value of good lives can always be outweighed by the disvalue of sufficiently many lives with negative lifetime welfare levels. More exactly, for any background population and any population, X, consisting entirely of lives with positive welfare levels, it holds that, given any negative welfare level, there is some number, n, such that the addition of at least n individuals at that negative level alongside X makes their joint addition to the background population morally undesirable.

Call this principle *Always Outweighable* (AO). VTU falsifies AO, presuming that we allow the vector elements to span negative as well as positive real values. Consider a null population, ∅, in which nothing of value exists. Letting the value of ∅ be represented by the vector (0, 0), VTU entails that for some[10] $\pi > 0$, any ρ, and any $\sigma < 0$, a population of m people at (π, ρ) and n people at $(0, \sigma)$ is better than ∅ for any n. There are some lives so good that their addition to the population can justify the addition of any number of lives that are only barely not worth living.

The falsity of AO bears in its own way on the desirability of the continued existence of human beings. If there are some lives so good that their addition to the population can justify the addition of any number of lives that are only barely not worth living, it may not seem altogether unreasonable to suppose that some human lives stand in this relationship to a significant proportion of intensively farmed animals. Compare, in particular, the best human lives against lives whose welfare level is like that of the average American broiler chicken (see Norwood and Lusk 2011: 128–131).

The vast majority of terrestrial vertebrates slaughtered for food are chickens. In 2007, nearly 9 billion broilers were produced in the US. Many believe that broilers have lives that are not worth living. Herzog (2010: 155) describes the conditions in which broilers are raised as 'Dante-esque: the chicks will never see sun nor sky. Because they are so top-heavy, broiler chickens spend most of their day lying down, often in litter contaminated with excrement. As a result, many will develop breast blisters, hock burns, and sores on their feet.' However, Norwood and Lusk (2011) reject the view that broilers have lives not worth living. While conceding that 'broilers lead short, unexciting lives', (131) they conclude that 'after reviewing all the obstacles to welfare and the nature of the birds, in our assessment, broiler farms do not cause large-scale suffering' (131). They note that pre-slaughter mortality rates are low among broilers, and lower than most people expect, at just 5%. Furthermore, the birds have ample food and water, as well as litter for scratching.

Someone might not unreasonably conclude that if broiler chickens on average have lives that are not worth living, their lives might nonetheless be only barely not worth living. Since the idea of a life that is 'only barely not worth living' is so unclear,[11] there is

[10] Given the lexicographic ordering, this is true for any $\pi > 0$. Relative to the ordering described in the preceding footnote, it is true for any π, m such that $\pi > \Delta / m$.

[11] As, of course, is the idea of a life that is 'barely worth living'; see Cowie (2017).

little point in disputing this exact form of words.[12] The key point is this. If we reject AO, it seems open to us to claim that the best human lives are such that their value cannot be outweighed by any number of lives at a negative welfare level corresponding to a not-too-absurd estimate of the average lifetime welfare of the American broiler chicken. This opens the door to a novel way of resisting some arguments for the claim that human extinction may be morally desirable.

5 Critical-level utilitarianism

In this section, I consider a generalization of TU, known as *critical-level utilitarianism* (CLU). According to CLU, x is at least as good as y if and only if

$$\sum_{i=1}^{N(x)}(u_i(x)-c) \geq \sum_{i=1}^{N(y)}(u_i(y)-c)$$

In other words, we compare the sum of each person's welfare net of c, and regard the first outcome as at least as good as the second just in case this quantity is at least as high in the first outcome as in the second. Here, c is the so-called *critical level*: the lifetime welfare level at which the addition of a person to the population is exactly as good as the outcome in which that additional person never exists. If $c = 0$, CLU reduces to TU. We might instead prefer to set c to be positive (Broome 2004; Blackorby, Bossert, and Donaldson 2005). Call this view PCLU.

Although PCLU rules out PN, it behaves in a sense like a weak form of PA, at least in the following respect: it entails that for any two welfare levels equidistant from zero, adding one life at the positive and one at the negative level is worse overall than adding neither. In that sense, good lives are not as good as bad lives are bad. However, it is also morally bad to add a person at the zero level, and at any positive welfare level less than c, unlike on PA.

I will also say a little bit about why we might accept PCLU. On the one hand, PCLU can be motivated by a desire to avoid RC. PCLU has also been motivated by appeal to the fact that it can give weight to the value of *unfragmented lives* (Blackorby et al. 2005: 151–152) or *longevity* (Broome 2004: 257–259), since PCLU counts, say, a long life lived at level 10 as better than two proportionally shorter lives, each lived at level 5. As we'll now see, these motivations can come into conflict in a way that matters when assessing the bearing of PCLU on the value of the future.

The conflict can be brought out by noting that it is plausible that the vast majority of non-human animals, even granting that they have positive welfare levels, nonetheless have

[12] Granting that individual welfare levels are represented by a two-dimensional vector and that populations are ordered by VTU, someone might object that it would have to be a strange coincidence if the welfare levels of non-human animals of the kind discussed here just happened to receive a score of exactly 0 on the first dimension, as opposed to, say, 0.00000001. But that needn't seem like a strange coincidence at all if we're independently attracted to a view on which the most important components of well-being, such as autonomy and meaningfulness, don't admit of arbitrarily small values. See Nebel (2022: 207) for discussion of the plausibility of this view and its significance for the question of whether VTU avoids RC.

welfare levels below the critical level, at least when the critical level is chosen to satisfy our intuitions about RC (Williamson 2021). Some arguments supporting this conclusion were reviewed in the previous section. To these we may add the observation that many wild non-human animals adopt life-history strategies that emphasize high rates of reproduction and minimal parental investment, producing offspring in quantities several orders of magnitude greater than the replacement rate, almost none of which survive to reproduce. Some conclude, on this basis, that most wild animals have lives that are not worth living and that suffering predominates in nature (Ng 1995; Horta 2010; Tomasik 2015). These arguments are contestable (Cuddington 2019; Groff and Ng 2019; Browning and Veit 2023). Nonetheless, even if these lives are above the zero level, if PCLU is true, they presumably have negative contributory value.

On the other hand, Blackorby et al. (2005) explicitly argue for a species-variable critical level, justifying this in terms of the idea that a positive critical level is intended to give weight to the value of unfragmented lives. Plausibly, these values are instantiated only to the extent that lives are unified over time via a sense of oneself as a creature with a past and a future. Blackorby et al. (2005: 326) write: 'for species whose members have less integrated lives, a lower critical level is reasonable, with a level of zero for species whose members exist in the present only'. Call a view of this kind SV-CLU.

The problem is that SV-CLU entails RC, at least if we assume that welfare is a scalar quantity. Suppose that we apply a positive critical level to human lives but not to ants. Then SV-CLU entails that for any large population of humans at very positive welfare levels, there is a larger population of ants whose lives are barely worth living whose existence is better. If our motivation for accepting PCLU is to avoid RC, this gives us reason to reject SV-CLU. In that case, however, we should be reasonably confident that most non-human animals, even if they have positive lifetime welfare, have negative contributory value given PCLU, and that the world would be better without almost everything that moves in the waters, in the air, or upon the earth. The tendency of human civilization to defaunate the planet would no longer be much of a strike against us, nor would the desirability of our continued existence receive support from claims that we are the biosphere's only hope of surviving on timescales beyond a billion years.

It is not just non-human animals, of course. Possibly, many existing people have lives only barely worth living (see Tännsjö 2002). Looking to the future, consider the scenario described by Hanson (2016), in which the economy is made up principally of *emulated minds* ('ems') derived by scanning human brains. One striking feature of 'ems' is the ease with which they can be copied at scale. Hanson predicts that in an 'em' economy heavy use will be made of spurs, which are very short-lived copies existing for periods of minutes or hours, used to accomplish one-off tasks, such as search tasks that allow for fast parallel exploration of a search space. Hanson (2016: 195) expects 'most em work to be done by spurs'. While the contours of personal identity become fuzzy in this scenario, it seems plausible that spurs, purely by virtue of being so extraordinarily short-lived, are individuals with lifetime welfare levels below c when c is set at a positive value capable of satisfying our intuitions about RC.[13] Hanson's projected future therefore looks to be a moral catastrophe given PCLU.

[13] In other words, a world of spurs constitutes a *Short-Lived Z* outcome in the sense defined by Portmore (1999).

6 Average utilitarianism and variable value theories

I know of no one who treats the promotion of happiness as fundamentally more important than the alleviation of suffering. However, there are some fairly intuitive population axiologies that can behave in practice like CLU with a negative critical level. Such views may in practice count the addition of good lives as making the outcome better than it would be made worse by the addition of a life not worth living equidistant from the zero level, *ceteris paribus*.

Specifically, we will look at *average utilitarianism* (AU) and theories that behave like AU in the large population limit. On AU, x is at least as good as y just in case

$$\frac{1}{N(x)}\sum_{i=1}^{N(x)}u_i(x) \geq \frac{1}{N(y)}\sum_{i=1}^{N(y)}u_i(y)$$

In other words, we compare the sum of welfare in each population, divided by the population size, and regard the first outcome as at least as good as the second just in case this quantity is at least as high in the first outcome as in the second. According to *variable value theory* (VVT) (Hurka 1982; Ng 1989), we instead have that x is at least as good as y just in case

$$\frac{g\big(N(x)\big)}{N(x)}\sum_{i=1}^{N(x)}u_i(x) \geq \frac{g\big(N(y)\big)}{N(y)}\sum_{i=1}^{N(y)}u_i(y)$$

where $g(\bullet)$ is a strictly concave, strictly increasing function with a horizontal asymptote. Thus, we now compare the average welfare level, weighted by a function of the population size whose slope decreases as the population gets bigger and bigger and eventually ends up approximately flat. This function is chosen to ensure that VVT agrees approximately with AU in evaluating population changes when the background population is large, and approximately with TU when the background population is small. By 'the background population', I mean those individuals whose existence at a given welfare level is independent of the possible population changes under evaluation.

As Tarsney and Thomas (2020) show, given a suitably large background population, AU and VVT behave approximately like CLU in the evaluation of population changes, when c is set to be the average welfare of the background population. Therefore, if the average welfare of the background population is negative, AU and VVT behave approximately like CLU with a negative critical level.

As Tarsney and Thomas (2020: 22–23) note, there is in fact a case to be made that there exists a large background population with negative average welfare. That is because the vast majority of the historical terrestrial population is composed of wild animals. While the truth of the matter is difficult to judge, there are good reasons to fear that wild non-human animals have negative lifetime welfare on average, for reasons noted previously.

Of course, even if wild animals on Earth have negative average lifetime welfare, the issue remains open, since the terrestrial biosphere may be only a small fraction of the true background population. On the other hand, it may be thought that if there is life elsewhere, then, absent

evidence to the contrary, the expectation of its average welfare level should not differ from the average for terrestrial life. However, the average welfare level for Earth-originating life over all time need not be the same as the average welfare level of the terrestrial background population. Keep in mind that the background population includes all those individuals whose existence at a given welfare level is independent of the possible population changes under evaluation. For our purposes, the future terrestrial population is not part of the background population, whereas developments paralleling those that may occur in our future form part of the background population for all suitably distant planets. Thus, setting our credences such that we do not expect the average welfare level elsewhere in the universe to differ from the terrestrial average need not mean that we do not expect the average to differ from the terrestrial background population.

The future, after all, is potentially enormous. The extinction of terrestrial multicellular life is projected at 0.8–1.3 billion years from today (Franck, Bounama, and von Bloh 2006). Space settlement, astroengineering, and orbital changes could allow for the survival of Earth-originating welfare subjects far beyond that (Sandberg forthcoming). As members of the only terrestrial genus with a capacity for cumulative technological culture, the future of Earth-originating life could be dominated by our descendants.

On the other hand, we may be especially wary of projecting terrestrial outcomes associated with the existence of organisms with a cumulative technological culture to exobiospheres, in light of the Fermi Paradox (Webb 2002). The observable universe is projected to contain roughly 10^{20} Earth-like planets, with 10^9 in the Milky Way alone; the Earth is a late-comer among them, being among the last 20% of Earth-like planets to form (Behroozi and Peeples 2015).[14] Nonetheless, we do not find evidence of life elsewhere in the universe. Arguably, this observation requires us to reject the Copernican assumption that the Earth is typical among Earth-like planets. One possibility is that it is atypical in respect of the emergence of organisms with a cumulative technological culture against the background of a biosphere populated by complex animals, whose existence is itself not especially improbable (Powell 2020: 267–278).

7 Prioritarianism and egalitarianism

Each of the utilitarian axiologies discussed in sections 4–6 is indifferent to how a fixed sum of welfare is distributed when the population is fixed. In that sense, they are insensitive to the moral significance of distributive patterns.

An alternative theory that answers to this concern is *prioritarianism* (Parfit 1991; Holtug 2010; Adler 2011). Roughly, this is the view that improvements to a person's welfare are of greater moral value if the person's welfare level is of a lower absolute level. For concreteness, I shall interpret prioritarianism as *total prioritarianism* (TP), the view that x is at least as good as y just in case

$$\sum_{i=1}^{N(x)} f(u_i(x)) \geq \sum_{i=1}^{N(y)} f(u_i(y))$$

[14] 'Earth-like' here means having an orbital radius and energy flux similar to that of Earth, allowing for stable surface reservoirs of liquid water.

where f is a strictly monotone increasing, strictly concave function with zero as a fixed point. Thus, we compare the sum of a function of each person's welfare, where the function doesn't increase linearly as a person's welfare improves, its slope instead decreasing more and more as a given individual is better and better off.

Note that, since f is strictly concave, it is worse to worsen lives that are already not worth living than it is good to improve lives that are already worth living by the same amount. Furthermore, for two welfare levels equidistant from zero, it is worse to add a person at the negative level than it is good to add the person at the positive level. In that sense, lives that are not worth living are bad to a greater degree than lives that are worth living are good (Holtug 2010: 255–256). Thus, prioritarianism supports a weak form of PA and a more pessimistic attitude toward the value of the future than a view like TU, which weights good and bad lives symmetrically. It bears some similarity to the view that 'suffering is bad to a greater degree than happiness is good' (Mayerfeld 1999: 158).[15]

Prioritarianism's main rival among distribution-sensitive theories is *egalitarianism*. The view on which equality is a feature of outcomes that contributes to their ranking as morally better or worse is the view Parfit calls *telic egalitarianism*. On the standard telic egalitarian view, it is in itself bad if some people are worse off than others through no choice or fault of their own (Temkin 1993; Segall 2016).[16] For concreteness, I assume that the currency of equality is lifetime welfare and that its scope is unrestricted in space and time. My focus will be on the plausibility of the following *egalitarian argument for anti-natalism* (EAAN), which I take to be orthogonal to these assumptions.

Here is an informal statement of EAAN. Inequality is in itself bad. If people continue to be born, there will be many more inequalities. Thus, things would be best with respect to equality if no more people are born (compare Temkin 1993: 216–217; Segall 2019: 421–422).[17] How plausible is this argument?

EAAN is most naturally read as assuming that the disvalue of inequality is increasing in the sum of the pairwise differences between people's welfare levels. This ensures that it will inevitably be worse from the perspective of equality if more people come into existence, since it is inevitable that more people will be worse off than others through no fault or choice of their own. Most simply, we might measure the level of inequality in an outcome, x, containing $N(x)$ persons, as

$$\frac{1}{2}\sum_{i=1}^{N(x)}\sum_{j=1}^{N(x)}\left|u_i(x)-u_j(x)\right|$$

In turn, the argument can be challenged by appeal to aggregate measures of the disvalue of inequality on which inequality is a function of the average size of the differences in people's welfare levels, since an increase in the sum of welfare differences due to an

[15] See Mayerfeld (1999:149–158) and Hurka (2010) for further discussion.

[16] For the sake of brevity, I treat the qualifier 'through no choice or fault of their own' as implicit from here on.

[17] Obviously, that is not to say that things would be best *all things considered*, since telic egalitarianism is compatible with a pluralist axiology on which the value of outcomes depends on, say, the level of inequality *and* the level of total welfare.

increasing population may be offset by the fact that we divide this sum by a larger population size. As an example, we may consider the *Gini Coefficient* (GC), which divides the sum of the pairwise differences between people's welfare levels by the square of the population size and by the average welfare level, $\overline{u}(x)$:

$$\frac{1}{2N(x)^2\,\overline{u}(x)}\sum_{i=1}^{N(x)}\sum_{j=1}^{N(x)}\left|u_i(x)-u_j(x)\right|$$

If the aggregate disvalue of inequality is measured by (7) rather than (6), then it is possible to improve the outcome with respect to inequality as a result of an increase in the population.[18]

It is also possible to challenge EAAN while retaining a simple additive measure of the badness of inequality across populations. The proper way of valuing equality, we might claim, should assign positive value to additional instances of pairwise equality between persons, and not merely disvalue to instances of pairwise inequality. Thus, according to Arrhenius (2013: 85) 'the more people who are unequal, the worse [is a population] in regards to egalitarian concerns, other things being equal; *and* the more people who are equal, the better [is a population] in regards to egalitarian concerns, other things being equal'. Arrhenius calls this *positive egalitarianism* (PE), in contrast with *negative egalitarianism* (NE), which is the view presupposed by EAAN. He notes that PE may allow us to count a two-person population in which one person is at 10 and another is at 20 as worse than a population in which one person is at 10 and 999 people are at 20, since, while the latter contains 999 relations of inequality by comparison with former's 1, it also contains 498,501 relations of equality, rather than none.[19]

We conclude that two key assumptions are required by EAAN. First, EAAN presupposes an additive measure of the badness of inequality. Second, EAAN presupposes NE. I do not have space to address the plausibility of these assumptions (see Temkin 1993: 191–231; Persson 2001; 2003; Rabinowicz 2003; Arrhenius 2013; Segall 2016; 2019; Gustafsson 2020; Arrhenius and Mosquera 2022). Nonetheless, in my view, both are credible, insofar as telic egalitarianism is.

[18] A different way in which to challenge EAAN's reliance on a simple additive measure of the badness of inequality is by claiming that the disvalue of inequality is conditional on existence. More exactly, we say that x is worse than y in respect of equality in virtue of the inequality between i and j in x only if i and j exist in y. This is in line with the view adopted by Parfit (1984: 425), who denies that we make the outcome worse with respect to equality by adding someone with a life worth living who is worse off than already existing people through no fault or choice of their own.

[19] We might also propose to resist EEAN by counting additional instances of pairwise equality between persons as neutral and as exhibiting 'greedy' neutrality of the kind discussed in section 3 of this chapter. However, arguments for attributing 'greediness' to the neutrality of additional good lives given PA do not seem to carry over to the value of equality. The argument that PA requires us to posit 'greedy' neutrality for good lives depends, ultimately, on the fact that, while the addition of a life worth living may be claimed to make the outcome neither better nor worse, any life can in principle always be better in respect of welfare, and its being so yields a morally better outcome, *ceteris paribus*. It is this structural fact about welfare and its value that requires us to reject the otherwise natural interpretation of PN as the claim that the addition of a life worth living leaves the outcome exactly equally as good as before (see Broome 2005: 405–407). Equality is importantly unlike welfare in this respect, in that some instance of pairwise equality between persons cannot be improved in respect of equality, in the way that any life worth living can be improved in respect of welfare.

8 Perfectionism

The axiologies discussed so far are all welfarist, in that they treat value as supervening on well-being. A prominent non-welfarist axiology is *perfectionism* (Hurka 1993). Very roughly, perfectionism attributes intrinsic value to the achievement of excellence in pursuits such as the arts and sciences, politics, or sport. Hurka (1993) develops a version of what he calls *narrow perfectionism*. In the tradition of Aristotle (350 BCE [2009]), Marx (1844 [2007]), Mill (1863), Nietzsche (1901 [2017]), and Nussbaum (2000), narrow perfectionism treats the good as the development to a high degree or full realization of those capacities that are central to human nature.

Perfectionism, in Hurka's conception, is not a theory of individual welfare. The perfectionist good is a distinct dimension of the good for which a person might choose to sacrifice her well-being.[20] One notable respect in which the narrow perfectionist good seems to differ from the good of well-being is by lacking an intrinsically bad, negative counterpart. With respect to a given person's welfare, outcomes can be classified as good, neutral, or bad. Pleasure has its counterpart in pain.[21] The narrow perfectionist good has been argued to have no intrinsically bad, negative counterpart, because there is no meaning to the idea of developing one's essentially human capacities to a negative degree (Hurka 1993: 100–101; Murphy 2001: 43–44; Sumner 2020: 429–431).

This might be thought to entail that narrow perfectionism cannot support a negative verdict on the value of the future. However, this does not follow. While the value of a sum cannot decrease as a result of the addition of non-negative terms, an average can decrease as a result of incorporating additional non-negative terms. Notably, Hurka (1993: 69–83) argues that counterparts of AU and VVT are more plausible as applied to perfectionist value, whereas a counterpart of TU is much less believable. When it comes to welfare, surely more of what is good is always better. The same does not obviously hold for excellence. Intuitively, some artistic or sporting careers would have been better had they ended sooner, not because their late stages do not contain (what are by ordinary standards) genuine achievements, but because they fall conspicuously beneath the level of excellence established by the artist or sportsperson earlier in their career. For example, Francis Ford Coppola's career as a filmmaker would arguably have been better overall in respect of the achievement of artistic excellence if he had retired to pursue winemaking already after completing *Apocalypse Now* in 1979.

For reasons already noted, if averaging (or averaging within the large population limit) is plausible as a principle for aggregating perfectionist value across people, then perfectionism can contribute to a negative assessment of the value of the future even without the postulate of intrinsic perfectionist bads. All that is required is that our descendants fall below the standard for perfectionist achievement set by previous generations. Futures in which perfectionist goods go into decline, sacrificed for the sake of comfort, ease, and safety, are a staple of cultural pessimist fears about modernity. Modernity supposedly makes it too

[20] Hurka (1993) actually leans toward, but stops short of fully embracing, a monistic *pure perfectionism*, on which the achievement of excellence is the only good.

[21] As noted by Kagan (2014) and Sumner (2020), it is much less obvious how to identify the negative counterpart of well-being relative to non-hedonistic theories of welfare. For example, it is not obvious how to define 'desire-frustration' such that desire-frustration can plausibly count as something bad in itself, rather than as the mere absence of a good, *viz*. desire-satisfaction.

easy to satisfy our wants and needs without effort or creativity, offers us an ever-expanding menu of mindless distractions, and encourages narrow individualism over the pursuit of shared, communal projects. The result is 'the loss of a heroic dimension to life' (Taylor 1991: 4), a world of 'secure and self-absorbed last men, devoid of . . . striving for higher goals in our pursuit of private comforts' (Fukuyama 1992: 328). This fear is taken to extremes in fictional dystopias like Huxley's *Brave New World* and Pixar's *WALL-E*.

There can be purely philosophical reasons to predict an eventual decline in the level of perfectionist value realized in a suitably long-lived humanity. According to what Hurka calls the *single-peak perfection principle*, the first instance of a given achievement, such as climbing a certain mountain or proving a certain theorem, is most valuable, and subsequent repetitions of the same achievement diminish in value to zero (Hurka 1993: 79–82). Unless the space of possible achievements is boundless, the single-peak perfection principle suggests that a decline in the value of humanity's achievements is inevitable in the long run.

However, the narrow perfectionist view also faces distinctive challenges in its application to the very long run. It understands excellence as the full realization of 'whatever properties are *essential to humans and conditioned on their being living things*'. (Hurka 1993: 16) Setting aside the concern that such properties are biologically dubious (Kitcher 1999), it is not at all clear how to apply this framework over long-run timescales where a fixed human nature cannot be assumed and humanity might conceivably transcend biology (Kurzweil 2005).

Notably, some of the declinist fears belonging to the genre of pessimism I have highlighted build in the worry that humans will 'degenerate' by becoming biologically adapted to stultifying aspects of modernity (Moynihan 2020: 312–322). H. G. Wells's time traveller explains the Eloi, the physically and intellectually diminutive people he encounters on the surface in the year 802,701, in these terms: 'Humanity had been strong, energetic, and intelligent, and had used all its abundant vitality to alter the conditions under which it lived. And now came the reaction of the altered conditions' (Wells 1895: 74). Narrow perfectionism seems to have the implication that, considered as a result of biological evolution under natural selection, the loss of all higher faculties from the genus *Homo* would not be regrettable, in that individuals so adapted do not fail in the realization of their nature in foregoing the use of the sophisticated cognitive capacities we think of as distinctive of our species. This feels hard to believe—and to square with the spirit of a perfectionist ethics.

Cases of this kind also indirectly serve to highlight a difficult problem that Hurka explicitly brackets: namely, how to account for members of other animal species within a perfectionist axiology. I have implicitly assumed that under narrow perfectionism, perfectionist goods are to be aggregated, in the first instance, within a given biological species. For example, in considering whether the average level of perfectionist goods will decline over the long term as a result of the supposedly stultifying effects of modernity on human beings, I have implicitly assumed that we are concerned with the average level of perfectionist value achieved by humans, rather than the relevant average being taken with respect to all living things. This arguably makes sense within the framework of narrow perfectionism, since it conceives of the good in each of its concrete instances as species-relative. The question remains how to aggregate across species to arrive at a global evaluation of outcomes.

9 Conservatism

According Cohen (2013: 149), (small 'c') conservatives 'exhibit a bias in favour of retaining what is of value, even in the face of replacing it by something of greater value'. Things which are of value are to be valued not merely as vehicles by which goodness or beauty enter the world, but for themselves. Therefore, we have reason to oppose their destruction, even when necessary to bring something of greater value into existence. Frick (2017) argues that our moral reasons for ensuring humanity's survival can be understood as conservative in nature. We ought to ensure our species' survival not because it is better for there to be more beings with good lives, but because humanity is valuable in itself in virtue of its unique intellectual, affective, and ethical capacities, and the appropriate response to recognizing the value of a thing involves caring about its continued existence.[22]

How might a conservative attitude toward the value of humanity shape our thinking about the value of the human future? First and foremost, let us consider the core point at which Frick (2017) is ultimately driving. If we accept PA, we cannot claim that the continued coming-into-being of new humans is desirable in light of the value of individual welfare considered as a good to be promoted. However, we can argue that the continued existence of humanity is desirable in virtue of the appropriateness of adopting a conservative attitude toward the value of humanity.

However, it is not so easy to reach the conclusion that the survival of humanity is all-things-considered desirable in this way, even if we adopt an optimistic forecast of future welfare levels. As we recall, PA entails PN, and the kind of neutrality attributed to good lives by PN is 'greedy'. Earlier, I illustrated 'greediness' in terms of the ability to swallow up bad things and neutralize them. But 'greediness' goes both ways. Adherents of PN also appear committed to the idea that adding lives worth living to the population can swallow up good things and neutralize them (Broome 2005: 409). It follows that if we accept PA, we have trouble arguing for the desirability of the survival of humanity even if we can point to values besides the promotion of welfare that speak in favour of a continued human presence, such as conserving existing things of value. The neutrality associated with the addition of worthwhile lives to the population threatens to swallow up and neutralize those other values, forcing us toward the conclusion that extinction would not be worse than continued human survival.

Here is a different issue to consider. Suppose that TU entails that the continued existence of humanity is undesirable. There will be slightly too much suffering. Any minimally plausible moral outlook must allow that if some suitably high proportion of future lives will be very horrible if we go on, then it can be better that humanity cease. Might a conservative claim that the importance of retaining existing things of value, even when they can be

[22] It may be objected that conservatism about value is not a hypothesis about value in the sense that AU or TU is a hypothesis about value and should therefore not form part of this inquiry. Cohen (2013: 155) even claims that conservatism is incompatible with maximizing consequentialism. His intuition, I take it, is that because conservatism says that we have reason to retain what is of value rather than replace it with something of greater value, it entails that we can have most reason to choose an outcome that is sub-maximal in the axiological ordering. This inference is invalid. The axiological framework assigns values to outcomes. Conservatism is a claim about the proper response to the value of individual things or individual people. It is an open question how to relate the ordering of outcomes in terms of moral value to assignments of moral value to particulars. In principle, there is no reason why we are barred from ordering possible worlds in such a way as to give weight to the longevity of valuable things.

replaced with something better, can nonetheless make the continued existence of humanity desirable in less extreme scenarios where TU entails that the continued existence of humanity is morally undesirable?

Plausibly not. In Cohen's discussion, the emphasis is on conserving not something which is intrinsically valuable *in some respect*, but something which is intrinsically valuable *on the whole*. Thus, Cohen insists that unjust social arrangements are not appropriate objects of the conservative attitude, even if they are valuable in some respect or other. Arguably, if the continued existence of humanity would yield so much suffering that TU recommends its extinction, this is strong evidence that humanity lacks intrinsic value in the overall sense Cohen has in mind.

We may have distinctive, valuable qualities, but that is not all. We are also capable of profound evil and of callous indifference to profound evil. Therefore, it is not obvious that humanity is an appropriate object of conservative valuing. Nozick (1989: 238–239) once argued that in the aftermath of the Holocaust, it 'now would not be a *special* tragedy if humankind ended . . . . That species, the one that committed *that*, has lost its worthy status.' The exact sense in which Nozick means to rule out that the end of humanity would be 'a *special* tragedy' is not altogether clear, but one plausible reading is that Nozick means to deny that it is fitting to value humanity in the distinctive way that conservatives might value the Grand Canyon or the paintings of Fra Angelico, as objects of value worth preserving in their own right.

Conservatism itself may be thought to yield distinctive reasons for finding humanity unworthy of being so valued. It is a natural moral framework within which to justify an imperative to conserve existing species and ecosystems against threats from human industrial civilization, and within which to lament all that has already been lost and is being lost. As a result of human activity, extinction rates for mammals, amphibians, birds, and reptiles over the past 500 years are at least as high as those associated with the previous 'Big Five' mass extinctions and could yield comparable extinction magnitudes within as little as 300 years (Barnosky et al. 2011). In the literature on existential risks from artificial intelligence, it is sometimes noted that the gulf between our intelligence and that of future software agents could place us in a position of vulnerability with respect to machine intelligence like that in which the rest of the biosphere stands to us (Bostrom 2014: vii; Russell 2019: 132–136; Ord 2020: 142–143). What is striking is the implication that, in respect of other living things, we are exactly the kind of existential threat the authors warn us of and hope to avert.

10 Conclusion

We have considered a number of different value theories discussed among contemporary moral philosophers and how they might bear on the value to be assigned to the continued survival of humanity and to efforts to reduce the risk of human extinction. The results, I think, are often surprising, sometimes disturbing, and occasionally hopeful. We have observed that NU does not so easily support the desirability of human extinction as is often alleged. Nor does PA speak for going gentle into that good night. The putative desirability of human extinction considered in relation to the deleterious effects of human activities on non-human animals invokes many classic problems in population axiology related to the valuation of lives that are only weakly positive, as well as some new ones related to the

valuation of lives that are only weakly negative. There is a plausible case to be made that telic egalitarians must inevitably look upon continued human survival as less desirable, that perfectionism counsels us to someday burn out rather than fade away, and, finally, that the conservative case for preserving humanity has been overstated.

Acknowledgements

For comments on previous drafts of the material of this chapter, I'm extremely grateful to Hilary Greaves, as well as to Teru Thomas and Tim L. Williamson, and to Jacob Barrett, Tomi Francis, Christian Tarsney, and David Thorstad.

References

Adler, M. (2011), *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis* (Oxford University Press).
Aristotle. (350 bce [2009]), *The Nicomachean Ethics*, ed. L. Brown, transl. D. Ross (Oxford University Press).
Arrhenius, G. (2000) 'An Impossibility Theorem for Welfarist Axiologies', in *Economics and Philosophy* 16/2: 247–266.
Arrhenius, G. (2013), 'Egalitarian Concerns and Population Change', in N. Eyal, S. A. Hurst, O. F. Norheim, and D. Wikler (eds.), *Inequalities in Health: Concepts, Measures, and Ethics* (Oxford University Press), 74–92.
Arrhenius, G. and Mosquera, J. (2022) 'Positive Egalitarianism Reconsidered', in *Utilitas* 34/1: 19–38.
Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., and Ferrer, E. A. (2011), 'Has The Earth's Sixth Mass Extinction Already Arrived?', in *Nature* 471: 51–57.
Behroozi, P. and Peeples, M. S. (2015), 'On the History and Future of Cosmic Planet Formation', in *Monthly Notices of the Royal Astronomical Society* 454/2: 1811–1817.
Benatar, D. (2006), *Better Never To Have Been: The Harm of Coming Into Existence* (Oxford University Press).
Blackorby, C., Bossert, W., and Donaldson, D. J. (2005), *Population Issues in Social Choice Theory, Welfare Economics, and Ethics* (Cambridge University Press).
Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).
Broome, J. (2004), *Weighing Lives* (Oxford University Press).
Broome, J. (2005), 'Should We Value Population?', in *Journal of Political Philosophy* 13,4: 399–413.
Browning, H. and Veit, W. (2023), 'Positive Wild Animal Welfare', in *Biology & Philosophy* 38: 14.
Carlson, E. (1998), 'Mere Addition and Two Trilemmas of Population Ethics', in *Economics and Philosophy* 14/2: 283–306.
Carlson, E. (2007), 'Higher Values and Non-Archimedean Additivity', in *Theoria* 73/1: 3–27.
Cohen, G. A. (2013), 'Rescuing Conservatism: A Defense of Existing Value (All Souls version)', in G. A. Cohen, *Finding Oneself in the Other* (Princeton University Press), 143–174.
Cowie, C. (2017), 'Does the Repugnant Conclusion Have Any Probative Force?' in *Philosophical Studies* 174: 3021–3039.
Cuddington, K. (2019), 'Insect Herbivores, Life History and Wild Animal Welfare', https://rethinkpriorities.org/publications/insect-herbivores-life-history-and-wild-animal-welfare. Last accessed 10 January 2025.
Dawkins, R. (1995), *River Out of Eden: A Darwinian View of Life* (Basic Books).
Dirzo, R., Ceballos, G., and Ehrlich, P. R. (2022), 'Circling the Drain: The Extinction Crisis and the Future of Humanity', in *Philosophical Transactions of the Royal Society B* 377: 20210378.
Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J. B., and Collen, B. (2014), 'Defaunation in the Anthropocene', in *Science* 345/6195: 401–406.
Franck, S., Bounama, C., and von Bloh, W. (2006), 'Causes and Timing of Future Biosphere Extinctions', in *Biogeosciences* 3/1: 85–92.
Frick, J. (2017), 'On the Survival of Humanity', in *Canadian Journal of Philosophy* 47/2–3: 344–367.
Fukuyama, F. (1992), *The End of History and the Last Man* (Free Press).

Greaves, H. (2017), 'Population Axiology', in *Philosophy Compass* 12/11: e12442.
Groff, Z. and Ng, Y.-K. (2019), 'Does Suffering Dominate Enjoyment in the Animal Kingdom? An Update to Welfare Biology', in *Biology and Philosophy* 34/4: 40.
Gustafsson, J. (2020), 'The Levelling-Down Objection and the Additive Measure of the Badness of Inequality', in *Economics and Philosophy* 36/3: 401–406.
Hanson, R. (2016), *The Age of Em: Work, Love, and Life When Robots Rule the Earth* (Oxford University Press).
Herzog, H. (2010), *Some We Love, Some We Hate, Some We Eat: Why it's So Hard to Think Straight About Animals* (HarperCollins).
Holtug, N. (2004), 'Person-Affecting Moralities', in J. Ryberg and T. Tännsjö (eds.), *The Repugnant Conclusion: Essays on Population Ethics* (Kluwer), 129–161.
Holtug, N. (2010), *Persons, Interests, and Justice* (Oxford University Press).
Horta, O. (2010), 'Debunking the Idyllic View of Natural Processes: Population Dynamics and Suffering in the Wild', in *Telos* 17/1: 73–90.
Hurka, T. (1982), 'Value and Population Size', in *Ethics* 93/3: 496–507.
Hurka, T. (1993), *Perfectionism* (Oxford University Press).
Hurka, T. (2010), 'Asymmetries in Value', in *Noûs* 44/2: 199–223.
Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., Puranen, B., et al. (eds.) (2014), *World Values Survey: Round Six - Country-Pooled Datafile Version* (JD Systems Institute), https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Last accessed 10 January 2025.
Kagan, S. (2014), 'An Introduction to Ill-Being', in *Oxford Studies in Normative Ethics* 4: 261–288.
Kitcher, P. (1999), 'Essence and Perfection', in *Ethics* 110/1: 59–83.
Kitcher, P. (2000), 'Parfit's Puzzle', in *Noûs* 34/4: 550–577.
Knutsson, S. (2019), 'The World Destruction Argument', in *Inquiry* 64/10: 1004–1023.
Kurzweil, R. (2005), *The Singularity Is Near: When Humans Transcend Biology* (Viking).
Marx, K. (1844 [2007]), *Economic and Philosophical Manuscripts of 1844*, transl. M. Milligan (Dover).
May, T. (2018), 'Would Human Extinction Be a Tragedy?', in *The New York Times*, 17 December 2018, https://www.nytimes.com/2018/12/17/opinion/human-extinction-climate-change.html. Last accessed 10 January 2025.
Mayerfeld, J. (1999), *Suffering and Moral Responsibility* (Oxford University Press).
McCauley, D. J., Pinsky, M. L., Palumbi, S. R., Estes, J. A., Joyce, F. H., and Warner, R. R. (2015), 'Marine Defaunation: Animal Loss in the Global Ocean', in *Science* 347/6219: 1255641–1–1255641–7.
McMahan, J. (1981), 'Problems of Population Theory', in *Ethics* 92/1: 96–127.
Mill, J. S. (1863), *Utilitarianism* (Parker, Son, and Bourn).
Moynihan, T. (2020), *X-Risk: How Humanity Discovered Its Own Extinction* (Urbanomic).
Murphy, M. C. (2001), *Natural Law and Practical Rationality* (Cambridge University Press).
Nebel, J. M. (2019), 'Asymmetries in the Value of Existence', in *Philosophical Perspectives* 33/1: 126–145.
Nebel, J. M. (2022), 'Totalism Without Repugnance', in J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan (eds.), *Ethics and Existence: The Legacy of Derek Parfit* (Oxford University Press), 200–231.
Ng, Y.-K. (1989), 'What Should We Do About Future Generations? Impossibility of Parfit's Theory X', in *Economics and Philosophy* 5/2: 235–253.
Ng, Y.-K. (1995), 'Towards Welfare Biology: Evolutionary Economics of Animal Consciousness and Suffering', in *Biology and Philosophy* 10: 255–285.
Nietzsche, F. (1901 [2017]), *The Will to Power*, transl. R. K. Hill and M. A. Scarpitti (Penguin Books).
Norwood, F. B. and Lusk, J. L. (2011), *Compassion, by the Pound: The Economics of Farm Animal Welfare* (Oxford University Press).
Nozick, R. (1989), *The Examined Life: Philosophical Meditations* (Simon & Schuster).
Nussbaum, M. C. (2000), *Women and Human Development: The Capabilities Approach* (Cambridge University Press).
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Parfit, D. (1984), *Reasons and Persons* (Oxford University Press).
Parfit, D. (1986), 'Overpopulation and the Quality of Life', in P. Singer (ed.) *Applied Ethics* (Oxford University Press), 145–164.
Parfit, D. (1991), *Equality or Priority? The Lindley Lecture* (University of Kansas).
Parfit, D. (2011), *On What Matters: Volume* 2 (Oxford University Press).
Parfit, D. (2016), 'Can We Avoid the Repugnant Conclusion?' in *Theoria* 82/2: 110–127.
Persson, I. (2001), 'Equality, Priority and Person-Affecting Value', in *Ethical Theory and Moral Practice* 4/1: 23–39.

Persson, I. (2003), 'The Badness of Unjust Inequality', in *Theoria* 69/1–2: 109–124.
Popper, K. R. (1945 [2011]), *The Open Society and Its Enemies* (Routledge).
Portmore, D. W. (1999), 'Does the Total Principle Have Any Repugnant Implications?', in *Ratio* 12/1: 80–98.
Powell, R. (2020), *Contingency and Convergence: Toward a Cosmic Biology of Body and Mind* (MIT Press).
Rabinowicz, W. (2003), 'The Size of Inequality and Its Badness: Some Reflections Around Temkin's *Inequality*', in *Theoria* 69/1–2: 60–84.
Rabinowicz, W. (2009), 'Broome and the Intuition of Neutrality', in *Philosophical Issues* 19: 389–411.
Rachels, S. (1998), 'Is It Good To Make Happy People?', in *Bioethics* 12/2: 93–110.
Roberts, M. A. (2011), 'An Asymmetry in the Ethics of Procreation', in *Philosophy Compass* 6/11: 765–776.
Russell, S. (2019), *Human Compatible: AI and the Problem of Control* (Allen Lane).
Sandberg, A. (forthcoming), *Grand Futures: Visions and Limits of What Can Be Achieved.*
Segall, S. (2016), *Why Inequality Matters: Luck Egalitarianism, Its Meaning and Value* (Cambridge University Press).
Segall, S. (2019), 'Why We Should Be Negative About Positive Egalitarianism', in *Utilitas* 31/4: 414–430.
Sikora, R. I. (1978), 'Is It Wrong to Prevent the Existence of Future Generations?', in R. I. Sikora and B. Barry (eds.), *Obligations to Future Generations* (White Horse Press), 112–166.
Smart, R. N. (1958), 'Negative Utilitarianism', in *Mind* 67/268: 542–543.
Sumner, W. (2020), 'The Worst Things in Life', in *Grazer Philosophische Studien* 97/3: 419–432.
Tännsjö, T. (2002), 'Why We Ought to Accept the Repugnant Conclusion', in *Utilitas* 14/3: 339–359.
Tarsney, C. and Thomas, T. (2020), 'Non-Additive Axiologies in Large Worlds', GPI Working Paper No. 9-2020 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/christian-tarsney-and-teruji-thomas-non-additive-axiologies-in-large-worlds/. Last accessed 10 January 2025.
Taylor, C. (1991), *The Ethics of Authenticity* (Harvard University Press).
Taylor, P. W. (1981), 'The Ethics of Respect for Nature', in *Environmental Ethics* 3/3: 197–218.
Temkin, L. S. (1993), *Inequality* (Oxford University Press).
Thomas, T. (2018), 'Some Possibilities in Population Axiology', in *Mind* 127/507: 807–832.
Tomasik, B. (2015), 'The Importance of Wild-Animal Suffering', in *Relations: Beyond Anthropocentrism* 3/2: 133–152.
Webb, S. (2002), *If the Universe Is Teeming with Aliens . . . Where Is Everybody? Fifty Solutions to Fermi's Paradox and the Problem of Extraterrestrial Life* (Copernicus Books).
Wells, H. G. (1895), *The Time Machine: An Invention* (H. Holt and Company).
Williamson, P. (2021), 'A New Argument Against Critical-Level Utilitarianism', in *Utilitas* 33/4: 399–416.

6

Longtermism in an Infinite World

Christian Tarsney and Hayden Wilkinson

1 Introduction

Longtermism is, very roughly, the thesis that the moral value of actions available to present-day agents is primarily determined by their potential effects on the far future.[1] The case for longtermism rests on the potentially vast scale of the future: Human-originating civilization could persist for millions or billions of years, and could spread across a large portion of the accessible universe, resulting in an enormous number of future people. If we can affect the welfare of those vastly many people (conditional on their existence), or if we can increase or decrease the probability that they come to exist, these effects might well have greater moral significance than the effects of our actions on the near future.

The most straightforward argument (but far from the only argument) for longtermism rests on an axiology—a theory of moral value for outcomes and risky prospects—that is *additive*, *impartial*, and *risk-neutral*.[2] Roughly, *additivity* means that the value of an outcome is a weighted sum of values realised at particular 'value locations' in that outcome (e.g., welfare realised in the lives of particular persons), and *impartiality* means that all locations receive the same weight in that sum (regardless of, for instance, their spatiotemporal location or relationship to a particular agent). These premises allow us to reason that, since the far future contains a potentially vast number of value locations (in particular, persons[3]), how things go in the far future potentially makes an enormous difference to the overall value of outcomes. *Risk neutrality* means that the value of a risky option is equal to the *expected value* of its outcome (i.e., a probability-weighted sum of the values of its various potential outcomes). This premise lets us reason that, even if we can only slightly affect the probabilities of a good vs. a bad long-term future for humanity, these small changes in probability can still be the primary determinant of the value of our actions, since the stakes are

[1] Note that this is a claim about what actions *would be best* (an *axiological* claim), not about what agents *ought* to do (a *deontic* claim). It is possible that we sometimes ought not perform the best available action—for instance, if you can prevent five murders by committing one, a deontologist might concede that *it would be better* if you committed murder (since it's better for one murder to occur rather than five), while still maintaining that you *ought not* commit murder. Thus, in focusing on longtermism as an axiological thesis, we are setting aside the question of whether what one *ought to do* is primarily determined by possible effects on the far future.

[2] Arguments for longtermism can also be made in the context of axiologies that are, for instance, *averageist*, *egalitarian*, *person-affecting*, and/or *risk-sensitive*, though with additional complications. For relevant discussion, see Pettigrew (2022), Buchak (2023), Thomas (2023), Tarsney & Thomas (2024), Greaves and MacAskill (this volume), Tarsney (forthcoming), Wilkinson (n.d.b.), among others. Much of what we say in this chapter generalises to arguments for longtermism based on these alternative axiologies, but we focus on risk-neutral totalism for simplicity.

[3] We use 'person' simply as shorthand for 'morally considerable welfare subject'.

Christian Tarsney and Hayden Wilkinson, *Longtermism in an Infinite World* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Christian Tarsney and Hayden Wilkinson 2025.
DOI: 10.1093/9780191979972.003.0006

so high. We will refer to the conjunction of these three principles as *risk-neutral totalism*.[4] (For a more precise definition, see section 3.)

It is also possible, however, that the future is not merely vast but *infinite*, containing infinitely many value locations and infinite total value and/or disvalue.[5] And while the potential *vastness* of the future suggests that it is extremely morally important how our actions affect the long-term future, it is much less clear that the potential *infinitude* of the future carries the same implication. In particular, the possibility of an infinite future threatens to undermine any case for longtermism based on risk-neutral totalism. First, if the future (or the universe as a whole) contains infinite value and/or disvalue, and if our actions are guaranteed to have only finite effects, then nothing we do can ever affect the total impartially weighted sum of value in the universe. This might be taken as a reason to reject either additivity or impartiality, since together they seem to imply, implausibly, that none of our actions matter. Or, if we stand by these principles and bite the bullet on their apparent nihilistic implication, it seems we must give up on longtermism and any other claims about what it's best to do. Second, the mere *possibility* of an infinite future implies that the *expected* total value of all our options is infinite or undefined. This similarly might lead us to reject at least one of additivity, impartiality, or risk neutrality, or alternatively to accept that these principles do not justify longtermism or any other substantive practical conclusion.[6]

There is a substantial body of research on the moral comparison of infinite worlds, in both philosophy and economics. Most proposals in this literature aim to extend additive theories of the value of outcomes from finite to infinite contexts—that is, to develop views that are additive in finite contexts while also delivering plausible verdicts in infinite contexts. Most, though not all, of these proposals also aim to retain some version of impartiality. And insofar as they consider risk (which many do not), the usual aim is similarly to extend *risk-neutral* theories from finite to infinite contexts. This literature has generated many sophisticated proposals, which show that it is possible to preserve the spirit of risk-neutral totalism while delivering at least some plausible verdicts in infinite contexts. Nevertheless, all such proposals have significant counter-intuitive implications. And indeed, there are various impossibility results showing that *any* axiology for infinite worlds (not just those consistent with risk-neutral totalism) must carry some counter-intuitive implications, and in particular must give up at least some of the *prima facie* attractive features that can be satisfied in finite contexts.

The challenges of infinite axiology thus threaten the case for longtermism in two ways. First, they might lead us to simply give up on risk-neutral totalism, in favour of moral views that are less favourable to longtermism. For instance, we might conclude that the only escape from these challenges is to abandon impartiality, or to abandon the project of axiology

[4] This label is convenient but potentially misleading: In the context of welfarist axiologies, the category of additive, impartial theories includes not just total utilitarianism but also critical-level and prioritarian axiologies. Risk-neutral totalism, as we are using the term, ranks risky options by their expected sum of value at particular locations, which need not be the same as the expected sum of welfare at particular locations, even if value is determined entirely by welfare.

[5] This is implied by the influential (though still disputed) inflationary paradigm in cosmology (see Knobe, Olum, and Vilenkin 2006: 50–51). It is also implied by at least some versions of the dominant flat-λ cosmological model, by which the universe will persist forever in a state that is capable of generating life through statistical fluctuations (see Carroll 2020: 11–16). By either view, for any local physical phenomenon, the universe will contain infinitely many near-perfect duplicates, with probability 1.

[6] It is important to note that the challenges of infinite axiology are not unique to risk-neutral totalism—very similar challenges arise for most axiologies, including the sorts of axiologies mentioned in fn. 2 above.

entirely in favour of a particularly extreme form of non-consequentialism that recognises no moral reasons to make the world better. Second, if we do find a satisfactory extension of risk-neutral totalism to infinite contexts, it might turn out that when we *apply* this extended view, accounting for the potential infinitude of our actual circumstances, practical conclusions like longtermism that seemed inescapable when we were assuming the world to be finite are no longer supported.

This chapter will consider to what extent the challenges of infinite axiology in fact threaten the case for longtermism—in particular, the case for longtermism based on risk-neutral totalism. Our conclusions will be tentatively positive for longtermism: First, while extant proposals for extending risk-neutral totalism to infinite contexts all face costs, those costs are not severe enough to scuttle the project entirely. Second, as we will show, most such proposals allow us, when we can only predictably affect a finite part of an infinite universe, to simply ignore the infinite unaffectable part of the universe and reason as if the finite affectable part were all that existed. Insofar as this is our actual situation, which it is to a good approximation, the risk-neutral-totalist case for longtermism can still go through even while accounting for the potential infinitude of the future. The possibility that our actions might have *infinite* predictable effects raises further challenges, but tends to strengthen the case for longtermism since those effects are almost certainly located in the far future, and any extension of risk-neutral totalism should regard them as overwhelmingly important.

We proceed as follows. Section 2 describes our formal framework. Section 3 introduces two minimal principles that are implied by almost all extant views in infinite axiology. Section 4 will consider the extent to which these principles allow us to rely on finite ethical reasoning of the sort employed in the risk-neutral-totalist case for longtermism, given the circumstances and choices we actually confront. Section 5 considers how the possibility that our choices have infinite predictable effects on the far future affects the case for longtermism. Section 6 considers whether the difficulties of infinite axiology force us to reject risk-neutral totalism, and what implications this might have for the case for longtermism. Section 7 sums up and highlights some especially important questions for future research.

2 Formal framework

Let's first introduce some terminology and notation (adapted from Wilkinson 2021a; 2023).

We assume, first, a domain O of *possible worlds* or *outcomes*. Each world contains some set of *value locations*, or simply *locations*, with which valuable events are associated. A location is a token entity of some common type that can exist (or have counterparts) across different outcomes. Locations might be persons, or person-stages, or positions in space and time, or something else.[7] Whatever locations are, there is an infinite set $\mathcal{L}$ of all possible locations. We assume that the value of an outcome is determined, in one way or another, by which locations exist, the value realised at each location, and perhaps other features of locations (e.g., their relative positions in time). And we assume that the value realised at each location can be represented by a real number, in a way that is order-preserving (i.e., greater numbers

[7] For a defence of adopting persons as the appropriate type, see Askell (2019). For arguments in favour of adopting spacetime positions, see Wilkinson (n.d.a) and Wilkinson (2021b).

correspond to greater degrees of value) and unique at least up to positive affine transformation (meaning that the numbers carry meaningful information about the relative size of *differences* in value).[8] Let $\mathcal{V} \subseteq \mathbb{R}$ represent the possible degrees of value that can be realised at locations. Then each outcome O_i determines a local value function $V_i : \mathcal{L} \to \mathcal{V} \cup \{\Omega\}$ that specifies the value realised at each location l in outcome O_i, with Ω representing the non-existence of the location in that outcome. The *total* value of an outcome is the sum of local value at all locations that exist in that outcome (formally, $\sum_{l \in \mathcal{L}: O_i(l) \neq \Omega} V_i(l)$), which we will abbreviate $Tot(O_i)$. This sum can, of course, be infinite or undefined.

We also wish to compare *prospects* (probability distributions) over outcomes, which correspond to the options from which real-world agents must choose under conditions of risk. The set of all possible prospects over outcomes is denoted by $\mathcal{P}$. For any prospect P_i, its probability of resulting in an outcome in set $\mathcal{O}'$ is given by $P_i(\mathcal{O}')$. We will abbreviate $P_i(\{O\})$ to $P_i(O)$ to denote the probability of a single outcome O. When a prospect results in some outcome O with probability 1, we allow O to denote the prospect as well as the outcome.

An *axiology* is an evaluative ranking of both outcomes and prospects. We assume that these two rankings must be consistent in the sense that one outcome is at least as good as another if and only if a prospect yielding the first outcome with probability 1 is at least as good as a prospect yielding the second with probability 1. Thus we use $\succcurlyeq$ (read, 'is at least as good as') to represent both the ranking of outcomes and the ranking of prospects. The relation $\succcurlyeq$ is a preorder: a binary relation that is reflexive and transitive, but not necessarily complete. As usual, $\succ$ is the asymmetric part of $\succcurlyeq$ (representing strict betterness) and $\sim$ is the symmetric part (representing equal goodness).

3 Two consensus principles

In this section we consider principles for extending risk-neutral totalism to infinite contexts. With the formal apparatus from above, risk-neutral totalism can be expressed as the following thesis:

> *Risk-Neutral Totalism*: For any outcomes O_i and O_j whose total values are finite, $O_i \succcurlyeq O_j$ if and only if its total value is at least as great.[9] Likewise, for any prospects P_i and P_j whose *expectations* of total value are finite, $P_i \succcurlyeq P_j$ if and only if its expectation is at least as great.[10]

We are looking for principles, then, that extend risk-neutral totalism in the sense of implying these biconditionals, while also implying at least some further comparisons in

[8] The latter assumption follows from additivity, which lets us make sense of the relative size of value differences. For instance, we can say that the difference in value realised at l_1 and l_2 is at least as great as that between l_3 and l_4 if and only if substituting l_1 for l_2 and l_4 for l_3 in an outcome will always yield an outcome that is at least as good as the original. Additivity guarantees that the size ordering of value differences, defined in this way, will be complete.

[9] Formally: If $Tot(O_i)$ and $Tot(O_j)$ are both finite, then $O_i \succcurlyeq O \Leftrightarrow Tot(O_i) \geq Tot(O_j)$.

[10] Formally: If $\mathbb{E}(Tot(P_i))$ and $\mathbb{E}(Tot(P_j))$ are both finite, then $P_i \succcurlyeq P_j \Leftrightarrow \mathbb{E}(Tot(P_i)) \geq \mathbb{E}(Tot(P_j))$, where $\mathbb{E}(Tot(P)) = \sum_{O \in \mathcal{O}} Tot(O)P(O)$.

cases where total value or its expectation are infinite or undefined. And, less formally, we also want a view that preserves the *spirit* of risk-neutral totalism—that is, the spirit of the underlying (albeit imprecisely stated) principles of additivity, impartiality, and risk neutrality.

Our foil in this search is a view we will call *naive* risk-neutral totalism.

> *Naive Risk-Neutral Totalism*: For any outcomes O_i and O_j whatsoever, $O_i \succcurlyeq O_j$ if and only if its total value is at least as great. Likewise, for any prospects P_i and P_j whatsoever, $P_i \succcurlyeq P_j$ if and only if its expected total value is at least as great.

This view is naive because it generalises risk-neutral totalism from finite to infinite contexts in a way that is simple and straightforward, but which a little reflection reveals to be implausible. Suppose, for instance, that in outcomes O_a and O_b exactly the same infinite set of locations exists, in the same spatiotemporal arrangement, but that each location has value 1 in O_a and 2 in O_b. Nearly everyone would agree that O_b is better than O_a, but naive risk-neutral totalism implies that they are equally good. More troublingly for practical purposes, naive risk-neutral totalism implies that if there is already infinite value and/or disvalue in the world, then no finite change (e.g., saving a life) can ever make things better or worse overall. Since the effects of the actions actually available to us appear to be finite, this strongly suggests that it doesn't matter what we do (at least from an axiological point of view), even in cases where it clearly *does* matter (e.g., when we have the opportunity to save a life).[11]

In search of a more plausible view, there have been many alternative proposals for extending the totalist ranking of outcomes and/or the risk-neutral totalist ranking of prospects to infinite contexts (e.g., Vallentyne 1993; Liedekerke and Lauwers 1997; Vallentyne and Kagan 1997; Bostrom 2011; Arntzenius 2014; Jonsson and Voorneveld 2018; Wilkinson 2021b; Clark n.d.). But, rather than describe these (often rather intricate) proposals in detail, we will examine a pair of uncontroversial principles that almost all of them uphold. As we will see, these principles by themselves go a long way toward rescuing risk-neutral-totalist reasoning from the threat of infinities.

The first of these principles we will call *Sum of Differences*. It says that we can compare two outcomes by summing up the *differences* in value at each value location, as long as this sum is well defined.

> *Sum of Differences (SoD)*: For any outcomes O_i and O_j, a sufficient condition for $O_i \succ O_j$ is that
>
> $$\sum_{l \in \mathcal{L}} (V_i(l) - V_j(l)) > 0$$

[11] Along with the challenge of evaluating outcomes with infinite or undefined value and prospects over those outcomes, risk-neutral totalism also faces a challenge evaluating prospects that have infinite or undefined expected value, even though all their possible outcomes have finite value. (Again, this challenge is not unique to risk-neutral totalism.) Such prospects include, for instance, the St. Petersburg game (Bernoulli 1738) and the Pasadena game (Nover and Hájek 2004). There are notable parallels between these two challenges, in theory (both involve trying to rank divergent sums) and in practice (both threaten to create widespread incomparability between our options, particularly in situations where our choices might affect the very far future). But in this chapter, to keep things manageable, we focus exclusively on the first challenge (of *outcomes* with infinite or undefined value).

either by converging unconditionally to a positive value, or by diverging unconditionally to $+\infty$, with $\Omega = 0$ (i.e., non-existence of a location is treated as equivalent to existence with value 0). Likewise, if this sum is equal to 0, then $O_i \sim O_j$.[12]

To illustrate, consider the following pair of outcomes. (In this array, columns represent possible locations and rows represent possible outcomes. Each number in the array gives the local value at a particular location in a particular outcome.)

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	…
O_1:	0	1	1	1	1	1	1	…
O_2:	1	0	0	1	1	1	1	…
$V_1 - V_2$:	−1	1	1	0	0	0	0	…

Naive risk-neutral totalism says that neither of these outcomes is better than the other, since the sum of values in each outcome is infinite. But Sum of Differences lets us compare O_1 and O_2 by summing the numbers in the bottom row, as long as that sum is well defined. In this case, since the sum is positive, we can conclude that O_1 is strictly better than O_2. Importantly for our purposes, in cases like this where two outcomes differ at only finitely many locations, Sum of Differences implies that we can equally well compare the two outcomes by comparing their subtotals of value *at just those locations where they differ*. Thus, from the fact that the subtotal value of O_1 from l_1 to l_3 is 2, and the corresponding subtotal for O_2 is 1, we can conclude that O_1 is strictly better than O_2.

While Sum of Differences compares only some pairs of infinite outcomes, the comparisons it does imply are all highly plausible, insofar as one finds additivity and impartiality plausible in finite contexts. Unsurprisingly, then, almost every proposal for extending impartial additive axiologies to infinite contexts implies Sum of Differences (with respect to its preferred kind of value locations, e.g., persons or spacetime positions).[13]

But Sum of Differences says nothing about how to compare prospects. For that purpose, we can extend it to a principle we will call *Sum of Value-Probability Differences* (SVPD). To state this principle, let $P(V(l) = v)$ denote prospect P's probability of yielding an outcome with value v at location l.[14]

Sum of Value-Probability Differences (SVPD): For any prospects P_i and P_j, a sufficient condition for $P_i \succ P_j$ is that

$$\sum_{(v,l)\in\mathcal{V}\times\mathcal{L}} v \times (P_i(V(l) = v) - P_j(V(l) = v)) > 0$$

[12] This principle is presented and defended by Vallentyne and Kagan (1997: 11), Lauwers and Vallentyne (2004: section 5), and Basu and Mitra (2007).

[13] The only exceptions we know of are the proposals of Liedekerke and Lauwers (1997), Bader (n.d.), and Clark (n.d.), which all violate the Pareto principle (see section 5) with respect to any possible kind of location.

[14] We assume for simplicity that prospects are discrete, and that the set $\mathcal{V} \subseteq \mathbb{R}$ of possible degrees of value at particular locations is countable.

either by converging unconditionally to a positive value or by diverging unconditionally to $+\infty$, with $\Omega = 0$ (i.e., non-existence of a location is treated as equivalent to existence with value 0). Likewise, if this sum is equal to 0, then $P_i \sim P_j$.

Informally, this principle tells us to consider, for each pair of a degree of value and a possible location, the difference in the probability of that degree of value being realised at that location if P_i is chosen vs. if P_j is chosen. We then multiply these probability differences by the degree of value concerned, and sum these terms across both locations and degrees of value to obtain an overall ranking of the prospects. In cases of only finitely many value locations and finite expected local value at each, SVPD agrees with risk-neutral totalism. And it can compare many infinitary prospects too, as we illustrate below. But, importantly, SVPD does not always yield a comparison—it only does so if the sum in the definition converges (or diverges to $+/-\infty$) unconditionally (i.e., regardless of the order in which the terms are summed).

The infinite axiology literature doesn't contain as many proposals for comparing prospects as it does for comparing outcomes. But every such proposal, if combined with SoD, implies SVPD.[15] Like SoD, then, SVPD is a relatively weak principle that should be mostly uncontroversial insofar as our goal is to extend risk-neutral totalism to infinite contexts.

To illustrate SVPD, consider the following pair of prospects. (As before, columns represent possible locations and rows represent possible outcomes. The probability of a particular outcome under a particular prospect is given in the first column.)

	$P_1(O_i)$		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	…
P_1	0.5	O_1:	2	2	0	1	1	1	1	1	1	…
	0.5	O_2:	2	0	0	0	0	0	0	0	0	…

	$P_2(O_i)$		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	…
P_2	0.5	O_3:	2	0	2	1	1	1	1	1	1	…
	0.5	O_4:	0	0	0	0	0	0	0	0	0	…

Here, P_1 and P_2 yield the same prospects for all locations except $l_1 - l_3$. Importantly, this need not imply that these locations are *unaffected* by the choice of prospect. For instance, perhaps the value realised at these locations depends on a fair coin flip, and choosing P_1 will cause these locations to have value 1 iff the coin lands heads, while choosing P_2 will cause them to have value 1 iff the coin lands tails. But the choice of prospect does not affect

[15] In fact, if combined with SoD, every such proposal strengthens SVPD by constraining the order of summation. The proposals of Bostrom (2011: 27–30), Arntzenius (2014: 55–56), and Meacham (2020) strengthen it to satisfy what could be called *Sum of Differences in Expectations*: that two prospects can be compared by first summing the relevant terms over all values of v at each location l, and only then summing over all locations l. Effectively, they perform the Sum of Differences over expectations at each location. On the other hand, Wilkinson (2023: 14) strengthens SVPD to satisfy what could be called *Expected Sum of Differences*: that two prospects can be compared by first summing those differences for the pairs (v,l) corresponding to each outcome, and only then summing over all outcomes. Effectively, the latter method takes the expectation of the Sum of Differences. And it turns out that these two stronger principles are incompatible: the first principle implies *ex ante* Pareto while the second implies statewise dominance; but, in infinite contexts, these two principles can conflict (Wilkinson 2023: section 4). Because SVPD requires unconditional convergence, it is neutral in these hard cases, and compatible with either *ex ante* Pareto or statewise dominance.

the *probability distribution* over local outcomes for any of these locations. So, intuitively, we would like to ignore them. Fortunately, SVPD lets us do this: for all $i > 3$ and all v, $P_1(V(l_i) = v) - P_2(V(l_i) = v) = 0$. This means that, in applying SVPD, we can simply ignore all these locations, and compare P_1 with P_2 by comparing only their differing expectations of local value at locations l_1 to l_3, as long as these are finite. Here, P_1 has an expected (subtotal) value of 3 across these locations, while P_1 has an expected value of 2, so SVPD tells us that $P_1 \succ P_2$. And, in general, when two prospects yield the same local prospects at all but finitely many locations, SVPD allows us to compare those prospects by comparing their expected subtotals at that finite set of locations.

4 Do the consensus principles let us ignore real-world infinities?

We now have two modest and plausible principles for comparing options in an infinite world. Each of them yields at least some verdicts in infinite cases that naive risk-neutral totalism can't handle satisfactorily. Specifically, we have seen that these principles let us compare pairs of outcomes (resp., prospects) in which local outcomes (resp., prospects) differ at only finitely many locations by ignoring all the locations where there's no difference and applying naive risk-neutral-totalist reasoning to the finite remainder, as if only those locations existed. In this section, we will consider whether this is enough to recover the real-world practical implications that we would expect from naive risk-neutral totalism if the world were finite, including in particular the risk-neutral-totalist case for longtermism as described earlier.

Whether we can indeed do so depends on what prospects we actually face, and so will depend in part on what kind of probability is morally relevant—the same real-world option might be associated with one distribution of objective chances over outcomes, a different distribution of epistemic probabilities, and yet another distribution of subjective credences. We will focus on epistemic probabilities, that is, the probabilities that it is epistemically rational for an agent to assign to particular outcomes on the supposition that she takes a given action. (But much of what we say will carry over to subjective credences.)

So, is our real-world situation such that the pairs of options we are required to evaluate beget different epistemic probability distributions over local value at only finitely many locations? The answer to this question depends on our empirical evidence concerning how much of the universe we can affect and in what ways. There is, on the one hand, substantial empirical reason to believe that, even if the universe is infinite, we can only affect a finite part of it. In particular, the impending heat death of the universe seems to promise an end to life as we know it, only finitely far in the future. And given that causal signals cannot travel faster than the speed of light, there is only finitely much room in our causal future for value to occupy, before heat death overtakes us.

On the other hand, there are various live hypotheses that do allow for infinite quantities of moral value in our causal future: For instance, some multiverse hypotheses (e.g., Smolin's 'cosmological natural selection' model; see Smolin 1992) imply that events in one 'universe' affect events in other universes, which would suggest that we can affect the moral value of events beyond the heat death of our local 'universe'. Other hypotheses suggest that a future civilisation may someday be able to perform infinite computations, potentially simulating an infinite number of minds, within the finite spatiotemporal limits of our pre-heat

death future light cone (Earman and Norton 1993; Tipler 1994). Finally, and perhaps most straightforwardly, some versions of the dominant cosmological model (called the flat-λ model) imply that morally valuable life will not cease upon heat death: individual brains, civilisations, and even galaxies will continue to be generated by random fluctuations, sometimes called *Boltzmann brains* or *Boltzmann universes* (Carroll 2020: 10). And the manner and timing of those fluctuations may be affected (albeit likely not in any predictable way) by our present actions—such fluctuations can be altered by even subtle changes in gravity (as by the Hawking effect) and electric field strength (as by the Casimir effect).

None of these hypotheses represents established physics and, in general, the claim that there are infinitely many value locations in our causal future seems on a much weaker footing than the claim that there are infinitely many value locations in the universe as a whole. Our own impression is that, of the hypotheses surveyed above, the Boltzmann brain hypothesis is by a significant margin the closest thing to a plausible implication of established physical theories.[16] And if the only source of infinite value and disvalue in our causal future is Boltzmann brains that arise by random fluctuations after the heat death of the universe, this seems to leave us in the happy condition where an infinite axiology satisfying SVPD will allow us to simply apply naive risk-neutral totalism to the part of the world we predictably affect: While our present choices may determine which Boltzmann brains come to exist and what experiences they have, the epistemic probability of any particular event after the heat death of the universe (e.g., a particular Boltzmann brain existing at a particular spacetime position) does not vary from option to option. At least, it is very hard to see how our evidence could distinguish our options in this way. Thus, any two options in present-day choice situations will yield the same local prospects for all possible locations after the heat death of the universe, and SVPD therefore allows us to simply ignore these locations.[17]

On the other hand, hypotheses on which our descendants may be able to intentionally create new universes (as in cosmological natural selection) or perform computational supertasks do allow us to *predictably* affect infinitely many locations (i.e., affect their prospects). For instance, by increasing the probability that humanity survives the coming century, we increase the probability that our descendants will someday deploy these technologies, and thereby increase the probability of existence for infinite numbers of potential persons. Similarly, attempts to change the institutions or future values of human-originating civilisation might increase or decrease the probability that a civilisation with these infinitary capacities would choose to use them.

Thus, it may be that the really difficult problems of infinite ethics arise in precisely those cases to which the longtermist thesis is supposed to apply—namely, choices that affect the epistemic probabilities of humanity's long-term survival or other important long-term outcomes (e.g., particular values prevailing in the far future). Arguably, many of our choices are not like this—for instance, your decision what to eat for breakfast may make no difference at all to the epistemic probability of humanity's long-term survival (for discussion,

[16] But, as Carroll (2020) explains, a universe eternally capable of generating Boltzmann brains is only implied by *some* versions of the flat-λ model, and we may well have reason to reject these versions exactly *because* they imply the existence of infinitely many future Boltzmann brains.

[17] Similar things can be said of certain multiverse hypotheses where we can affect other 'universes' (for instance, through gravitational interactions in a higher-dimensional space) but are not in a position to know anything about the empirical details of those effects.

see Greaves and Tarsney, this volume, section 3). In that case, minimal principles like SoD and SVPD can straightforwardly protect us from the paralysing effects of infinitary ethical considerations in these ordinary cases. But in the more consequential situations where our choices do have some predictable effect on the long-term future, they plausibly also make a non-zero—though perhaps *extremely* small—difference to the epistemic prospects of infinitely many potential value locations. So we must ask whether SVPD, or plausible extensions thereof, can handle these situations.[18]

5 Infinite predictable effects

Recall that the challenge of infinite axiology threatens the case for longtermism in two ways: (i) because it may force us to abandon risk-neutral totalism and with it the risk-neutral totalist case for longtermism (and similarly, despite our focus here, it may force us to abandon various other axiologies that support longtermism too); and (ii) because, if we do find a satisfactory way of extending risk-neutral totalism to infinite contexts, the practical implications of this extended view might deviate from what risk-neutral totalism would recommend if the universe were merely finite. The last two sections have gone some way toward mitigating both worries: We have seen that there are existing proposals for extending risk-neutral totalism that, in virtue of implying SoD and SVPD, can deliver at least some plausible verdicts in infinite contexts, rescuing us from universal infinitarian paralysis. And we have seen that when we can only affect the prospects of finitely many locations in an infinite universe, these principles yield the same practical implications that we would get by simply applying naive, finitary risk-neutral totalism to that finite part of the universe.

Nonetheless, both worries remain live. While extant proposals for extending risk-neutral totalism have some attractive features, they also have significant drawbacks, some of which (as we will see) are inescapable. And since we cannot rule out hypotheses that would allow us to predictably affect infinitely many value locations, it is not *quite* true that our actions only affect finitely many local prospects, so SVPD alone does not guarantee that the true infinite axiology will allow us to apply naive risk-neutral totalism as described above. In this section and the next we will consider these remaining worries, in reverse order.

First, then, suppose that (an extension of) risk-neutral totalism is true, and more specifically that SVPD is true. But suppose also that our choices do affect the local prospects of *infinitely* many locations, at least slightly. What practical implications does this have, particularly with respect to the case for longtermism?

This is a hard question to answer in general, partly because there are many importantly distinct ways in which our choices might affect infinitely many local prospects. But

[18] A different way in which our choices might affect the prospects of infinitely many locations is if the correct decision theory is non-causal (e.g., evidential). We have so far implicitly assumed a causal decision theory on which our choices can in principle only make a difference to the outcomes and prospects of locations in our causal future. But if evidential decision theory or some other non-causal decision theory is correct, then our options can yield different local prospects at locations outside our causal future (for instance, because our choices give us evidence about the choices of our doppelgängers in distant parts of the universe). If the universe is spatially infinite and contains infinitely many situations identical or arbitrarily similar to ours, then it is guaranteed that the number of such locations will be infinite as well. Indeed, even non-zero credence in non-causal decision theory might have this upshot, if we treat our uncertainty between causal and non-causal decision theory in the same way as empirical uncertainty (see MacAskill 2016; MacAskill et al. 2021). This would reinforce the conclusion in the main text that our choices may affect infinitely many local prospects, but perhaps only very slightly (if our credence in non-causal decision theories is only slight).

examination of a few particular cases will be enough to illustrate three general points: First, infinite predictable effects (i.e., affecting infinitely many local prospects) are not always problematic—in some cases, it is possible to rank pairs of prospects with this feature in a way that is principled, intuitively plausible, and in the spirit of risk-neutral totalism. Second, insofar as we can make comparisons in these situations, the possibility of infinite predictable effects will tend to *strengthen* the risk-neutral-totalist case for longtermism, since (i) risk-neutral totalists should generally give absolute priority to infinite effects over finite effects and (ii) these infinite effects will tend to be located in the far future. But, third, there are some kinds of infinite predictable effects, which we plausibly face in real-world choice situations, where it is intuitively unclear how to rank our options, where no ranking is given by modest principles like SVPD, and where it is at least conceivable that our options are simply incomparable. (Two options are *incomparable* if neither is better than the other, and they are not equally good.) The primary way in which infinite predictable effects might threaten the risk-neutral-totalist case for longtermism, then, is by implying that, in situations where our choices affect the long-term future, we face widespread incomparability, with no available option being better or worse than any other.[19]

To illustrate these points, let's start with the easy cases of infinite predictable effects, and work our way toward the harder cases.[20] First, there are cases of infinite predictable effects that SVPD ranks easily. For instance, suppose that there is some potential future population at infinitely many locations, each of whom will certainly have positive value if they exist, and you can increase the probability that they come to exist without changing their prospects conditional on existence.

$$P_3\begin{cases} P_3(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \dots \\ 0.5 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{cases}$$

$$P_4\begin{cases} P_4(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.51 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \dots \\ 0.49 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{cases}$$

In this case, the sum from the definition of SVPD—of $v \times (P_4(V(l) = v) - P_3(v(l) = v))$ for each location l and possible local value v—diverges unconditionally to $+\infty$ (bearing in mind that we treat Ω as 0). So, SVPD implies that $P_4 \succ P_3$.

There are other cases in which it seems clear which of two prospects is better, that are not ranked by SVPD, but that can be handled by natural and plausible strengthenings of SVPD. For instance, consider the following case, where you can improve the prospects of infinitely many locations conditional on existence, without affecting their probabilities of existence.

[19] While this conclusion might either refute longtermism or make it trivially true, depending on how the longtermist thesis is formulated, it would clearly violate the spirit of longtermism to conclude that we can never improve (the prospects of) the world as a whole by improving (the prospects of) the long-term future.

[20] We note in passing that effects on the outcomes or prospects of infinitely many locations can have finite sums—for instance, if we improve l_1 by $\frac{1}{2}$, l_2 by $\frac{1}{4}$, l_3 by $\frac{1}{8}$, and so on. These ultra-easy cases are handled adequately by SoD and SPVD, and in some cases by naive risk-neutral totalism. But if our choices do predictably affect infinitely many locations, the effects are unlikely to be this well-behaved. We focus, therefore, on situations involving infinite sums.

$$P_5\begin{cases} P_5(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.05 & O_2: & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \dots \\ 0.05 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \mathbf{1} & 1 & \dots \\ 0.9 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{cases}$$

$$P_6\begin{cases} P_6(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.051 & O_2: & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \dots \\ 0.049 & O_1: & 1 & 1 & 1 & 1 & 1 & \mathbf{1} & 1 & 1 & 1 & \dots \\ 0.9 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{cases}$$

Clearly P_6 is better than P_5. But SVPD is silent: Because P_6 increases each location's probability of realising value 2 while decreasing each location's probability of realising value 1, the value-weighted sum of location-outcome probability differences is non-convergent. But this can be remedied by strengthening SVPD, allowing outcomes at each location to be compared to a 'baseline' outcome for that location.

> *Baseline-Adjusted Sum of Value-Probability Differences*: For any prospects P_i and P_j, $P_i \succ P_j$ if there exists an outcome $O_b \in \mathcal{O}$ such that
>
> $$\sum_{(v,l)\in\mathcal{V}\times\mathcal{L}} (v - V_b(l))(P_i(V(l) = v) - P_j(V(l) = v)) > 0$$
>
> either by converging unconditionally to a positive value, or by diverging unconditionally to $+\infty$ (with $\Omega = 0$). Likewise, if there is an outcome O_b for which this sum is equal to 0, then $P_i \sim P_j$.

If we choose the outcome O_1 above (in which every location has value 1) as our baseline outcome O_b, and substitute P_6 and P_5 for P_i and P_j respectively, we find that the above sum diverges unconditionally to $+\infty$. So we can conclude that $P_6 \succ P_5$. And this baseline-adjusted principle, while slightly more complicated than SVPD, is similarly modest and uncontroversial.[21]

In both these cases, the principles we have appealed to imply that one prospect is 'infinitely better' than another in the sense that no finite improvement of the worse prospect (or worsening of the better prospect) could affect the comparison. (For instance, if we add any finite number of locations that will realise value 1 for sure under P_3 or P_5, and 0 for sure under P_4 or P_6, the ranking would be unchanged.) Correctly evaluating this sort of infinite improvement requires some extension of naive risk-neutral totalism.[22] But, in general, the possibility of such unambiguous infinite improvements *strengthens* the risk-neutral-totalist case for longtermism. Why? First, any infinite axiology in the spirit of risk-neutral totalism should be *fanatical* about infinite improvements: Shifting any amount of probability from

[21] In particular, like SVPD, it follows from the proposals in Bostrom (2011: 27–30), Arntzenius (2014: 55–56), Meacham (2020), and Wilkinson (2023: 353) given Sum of Differences.

[22] In both cases, the expected total value of both prospects is $+\infty$ (or undefined, if we do not countenance infinite expectations), so naive risk-neutral totalism rules that the two prospects are equally good (or simply fails to compare them).

an infinitely worse outcome to an infinitely better outcome should take precedence over any finitary considerations, in the evaluation of prospects (see Wilkinson 2022; Beckstead and Thomas 2024). And second, if there is any epistemic probability of our choices having such infinite effects, it is almost all in the far future: The infinitely-better and infinitely-worse trajectories whose probabilities we can affect will, presumably, either unfold over infinite future time, or require far-future technology (e.g., computers that can perform supertasks in finite time), or both.[23]

More generally, it seems to us that the possibility that we face choices between prospects that yield different local prospects at infinitely many locations does not threaten the case for longtermism *as long as these prospects can be compared*. As a rough argument: One version of the longtermist thesis is that our options typically differ more in far-future value than in near-future value. Suppose we believed this thesis while assuming that our choices only (predictably) affect finitely many value locations in the far future, but then come to believe that our choices affect infinitely many locations in the far future (without changing our beliefs about their effects on the near future). It seems unlikely (though not impossible) that this realisation should *reduce* the typical differences in far-future value between our options. This leaves two possibilities: One is that it amplifies those differences (or at least leaves them unchanged), thereby strengthening the case for longtermism (or at least leaving it unweakened). The other, however, is that we find that we can no longer compare the far-future effects of our options.

There are, unfortunately, many hard cases in infinite axiology that are not resolved by simple principles like SVPD, where it is not obvious how we should rank two outcomes or prospects, and where incomparability is plausible. Here are three examples. First, suppose your choice affects the probability that some infinite future population will come to exist (say, within an infinite simulation or a 'baby universe' of the sort envisioned by cosmological natural selection), and you know that if it does exist, that population will contain both infinitely many locations with positive value (e.g., persons with lives worth living) and infinitely many locations with negative value (e.g., persons with lives worth not living).

$$P_7\begin{cases} P_7(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & \dots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{cases}$$

$$P_8\begin{cases} P_8(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.51 & O_1: & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & \dots \\ 0.49 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{cases}$$

Second, suppose that your choice does not affect the probability that such an infinite future population comes to exist, but you believe it to be *identity-affecting*: that is, which particular locations will compose that population depends on your choice.

[23] Another conceivable source of infinite stakes that are not clearly located in the far future is supernatural— in particular, affecting the probabilities that particular individuals achieve infinitely good vs. infinitely bad afterlives. Whether these considerations count for or against longtermism depends on whether these possible afterlives are temporal, and whether they stand in temporal relations to the present.

$$P_9\begin{cases} P_9(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & 1 & \Omega & -1 & \Omega & 1 & \Omega & -1 & \Omega & 1 & \dots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{cases}$$

$$P_{10}\begin{cases} P_{10}(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & \Omega & 1 & \Omega & -1 & \Omega & 1 & \Omega & -1 & \Omega & \dots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{cases}$$

Third and finally, suppose you can affect the probability that this infinite future population will be governed by one set of norms, institutions, or values rather than another, for example, by having a potentially persistent effect on present-day values. For instance, you might affect the probability that future societies are inequality-averse, and thereby the probability of more versus less unequal distributions of welfare being realised in those societies.

$$P_{11}\begin{cases} P_{11}(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & 2 & 6 & 2 & 6 & 2 & 6 & 2 & 6 & 2 & \dots \\ 0.5 & O_2: & 3 & 4 & 3 & 4 & 3 & 4 & 3 & 4 & 3 & \dots \end{cases}$$

$$P_{12}\begin{cases} P_{12}(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.51 & O_1: & 2 & 6 & 2 & 6 & 2 & 6 & 2 & 6 & 2 & \dots \\ 0.49 & O_2: & 3 & 4 & 3 & 4 & 3 & 4 & 3 & 4 & 3 & \dots \end{cases}$$

None of these three cases are resolved by SVPD (or by the stronger, baseline-adjusted version discussed above). Nor is there an intuitively clear right answer in any of these cases.[24] That doesn't mean that these are necessarily genuine cases of incomparability—it's possible to articulate principles that deliver verdicts in cases like these, especially if those principles are allowed to take account of the spatiotemporal arrangement of locations (see, e.g., Wilkinson 2021b). But it is at least plausible that the kinds of trade-offs involved in these cases do create incomparability (especially since, as we will see in the next section, there are compelling formal arguments that there must be at least some incomparability in infinite axiology). And it is plausible that we face trade-offs like these in our real-world choices—at least, in those choices that have some predictable effect on the long-run future, which may affect the probabilities of infinite future populations coming to exist (e.g., by affecting the probability that our civilisation survives long enough to create them) or the prospects faced by those populations (e.g., by helping to shape the values and institutions that govern the far future).

We conclude then that, if the true axiology extends risk-neutral totalism, it will *either* leave the risk-neutral-totalist case for longtermism unscathed, *or* undermine it by implying widespread incomparability in real-world choice situations. Which of these possibilities is more plausible depends on at least three factors.

[24] In the second case, it is especially tempting to conclude that the two prospects are equally good, but natural ways of generalising this judgement can get us into trouble. For instance, as we will discuss in the next section, the principle of unrestricted anonymity (which says that two outcomes with the same cardinality of locations realising each degree of value are equally good) is incompatible with a weak Pareto principle and so, *a fortiori*, with SoD.

1. Our real-world epistemic situation—in particular, which hypotheses about the long-term effects of our actions we think deserve epistemic probabilities that are non-zero and non-symmetric (i.e., not cancelled out by equal probabilities of opposite effects, so that they create net differences between possible actions in the probabilities of particular long-term outcomes).
2. The strength of our infinite axiology—for instance, how often it is able to make comparisons between pairs of prospects where each is better at infinitely many locations, or where each is infinitely better in some states of nature. The former question might depend particularly on whether our axiology is sensitive to the spatiotemporal arrangement of locations, which can help us evaluate trade-offs between infinite sets of locations.
3. The criteria of identity or counterparthood for locations across different outcomes (which can, for instance, determine whether the choice between two prospects has the same effect at every location, or creates trade-offs between locations).[25]

Since all of these factors depend on difficult and unresolved philosophical questions, we unfortunately cannot yet decide with confidence between these two possible conclusions.

Many will judge, however, that *if* the most theoretically plausible extensions of risk-neutral totalism to infinite settings imply widespread incomparability in real-world choice situations, we should not embrace this conclusion but should rather abandon risk-neutral totalism.[26] For this reason, it seems that the most likely way in which the challenges of infinite axiology might undermine the risk-neutral-totalist case for longtermism is not by changing the practical implications of risk-neutral totalism, but by motivating its rejection. So let's next consider that possibility.

6 Giving up risk-neutral totalism

Why might the challenges of infinite axiology lead us to reject risk-neutral totalism (and its various possible extensions)? One important reason is the existence of impossibility results showing that some attractive features of risk-neutral totalism in finite contexts must be given up in infinite contexts. Depending on what principles one takes to be core commitments of risk-neutral totalism, these results might be taken to show that its core commitments are simply inconsistent, or that they are implausible since they conflict with other principles that, while not core commitments of risk-neutral totalism, are independently plausible. We will briefly mention four such results.

[25] For instance, suppose you face a choice that affects the probability that some future population containing both humans (with good lives) and non-human animals (with bad lives) will come to exist. Or suppose your choice influences the relative treatment of humans and non-human animals in the far future. If each possible location is either necessarily human or necessarily non-human, then these choices will involve trade-offs between locations—one option being better in expectation for some locations and worse in expectation for others. But if the identity or counterpart relation for locations is indifferent to species (as is plausible, for instance, if locations are spacetime positions rather than persons), then your choice might have the same or very similar effects on the local prospects of each location it affects, and so involve no trade-offs between locations. All else being equal, it will be easier to rank infinite prospects that do not involve trade-offs between two infinite sets of locations.

[26] In the literature on infinite axiology, the conclusion that *no* real-world option is better than any other is typically treated as a *reductio*, to be avoided at all costs. The exception is Smith (2003), who argues for *de facto* moral nihilism on the basis of broadly totalist moral assumptions coupled with the infinitude of the future.

The first is that even weak versions of the Pareto principle are incompatible with an unrestricted anonymity principle. *Weak Pareto* says that, if two worlds have the same locations but each location has greater value at one of the worlds, then that world is better. *Unrestricted Anonymity* says that if two worlds have the same locations, and have the same number (i.e., cardinality) of locations at each value level, then they are equally good—in other words, it doesn't matter *which* locations realise which degrees of value.[27] But both these principles seem to be core commitments of any theory that aims to extend risk-neutral totalism—Pareto is an extreme weakening of SoD, and Unrestricted Anonymity intuitively reflects the totalist commitment to impartiality.[28]

The second such result is that the (even Weak) Pareto principle cannot hold for different types of locations (at least for any two types of locations for which their counterpart relations are not essentially dependent on each other). For instance, it cannot be that an outcome is always made better by increasing the value obtained by each *person*, while also that an outcome is always made better by increasing the value obtained at each *position* in spacetime (Cain 1995; Wilkinson 2021b: 1925–1928). Again, risk-neutral totalism satisfies Pareto for all types of locations in the finite context, and each such version of Pareto may seem like a core commitment of totalist theories.

The third result concerns prospects. In the infinite setting, it cannot be true that both: increasing the expectation of value at every location always makes the prospect better (known as *ex ante (Weak) Pareto*); and replacing every outcome in a prospect with a better outcome (in the same state and with the same probability) always makes the prospect better.[29] Again, both principles are satisfied by risk-neutral totalism in the finite context and seem like core commitments of the view.[30]

The fourth and final result is that any ordering of infinite outcomes that satisfies both Weak Pareto and Finite Anonymity must be either incomplete or non-constructive.[31] Finite Anonymity says that permuting the local values at *finitely many* locations does not change the value of an outcome. This is much weaker than Unrestricted Anonymity, is consistent with Weak Pareto, and seems to straightforwardly reflect the ideal of impartiality. An ordering $\succcurlyeq$ is *non-constructive* if it does not have an explicit, finite description—that is, we cannot write down a true formula of the form 'For all O_i, O_j, $O_i \succ O_j$ if and only if φ', where the right-hand side of the formula does not contain $\succcurlyeq$ or anything defined in terms of it. While the viability of non-constructive axiological or normative principles has not been substantially explored (and we would find such exploration very valuable), it seems to us that any non-constructive axiology would be problematically arbitrary. Its specification would require infinitely many independent 'choices' to rank one outcome over another,

[27] The result comes from Liedekerke (1995) originally. See also Hamkins and Montero (2000: 237).

[28] It is contested, however, whether impartiality requires Unrestricted Anonymity—see Wilkinson (2021b: 1928–1931). In the literature, nearly all proposals to extend risk-neutral totalism opt to violate Unrestricted Anonymity to uphold Pareto (for at least some type of locations) and indeed SoD as well (e.g., Vallentyne 1993; Vallentyne and Kagan 1997; Jonsson and Voorneveld 2018; Wilkinson 2021b).

[29] This result comes from Wilkinson (2024: sec. 4); see also Hong and Russell (forthcoming).

[30] Indeed, stronger versions of both principles (restricted to a finite context) feature in the classic theorem of Harsanyi (1955) that is often taken to support risk-neutral utilitarianism.

[31] See Zame (2007: Theorem 4) and the more general result given in Lauwers (2010).

without any unifying principle to explain those choices—there would be an infinity of brute axiological facts.[32]

This fourth result, therefore, leaves us with a fairly strong argument for incompleteness. But this draws our attention to the second way in which infinities still threaten risk-neutral totalism: An infinite axiology that extends risk-neutral totalism, even if it satisfies all the theoretical desiderata we deem essential, may yield an unacceptable amount of incompleteness in practice. Apart from the axiomatic argument for incompleteness just described, and the cases in the last section where risk-neutral totalist commitments do not suggest any obvious ranking of alternative prospects, there are also general arguments for expecting fairly widespread incomparability in infinite extensions of risk-neutral totalism. For instance, it has been argued that any infinite axiology must generate widespread incomparability in practice if it satisfies Pareto for persons (Askell 2019; Wilkinson 2021b: sec. 3.1) or is insensitive to the spatiotemporal arrangement of persons, which is arguably a requirement of impartiality (see Wilkinson n.d.a.: secs. 3.2–3.4). Suppose that many of our choices turn out to have very small effects on the prospects of infinitely many potential value locations, with each option improving the prospects of infinitely many locations while worsening the prospects of infinitely many others, in such a way that our options are incomparable by risk-neutral totalist lights. Even if this is only true of some choices, and therefore does not leave us completely adrift in deciding what to do, it might nevertheless be seen as an unacceptable failure of the risk-neutral totalist worldview to offer practical guidance.

Suppose we conclude that these difficulties of extending risk-neutral totalism to infinite contexts are too great, and that risk-neutral totalism must therefore be given up. Importantly, the challenges of infinite axiology are not unique to risk-neutral totalism, and many ways of abandoning risk-neutral totalism would do little to ease these challenges—for instance, average utilitarian, prioritarian, and egalitarian views face similarly great difficulties. So what alternatives to risk-neutral totalism might we adopt if our main concern is to escape infinitarian difficulties altogether? Here are four possibilities.

1. **Pure time discounting**: Value and disvalue arising in the further future contributes less to our overall evaluation of outcomes merely because of its position in time. If our discount schedule is sufficiently severe (e.g., exponential) and value at a time is bounded, this implies that the total discounted value of the future is finite, even if the future contains infinitely many value locations.[33]
2. **Agent-relative consequentialism or strong non-consequentialism**: There is no such thing as the impartial or agent-neutral value of outcomes; or, if there is, it is largely

[32] For instance, one way of getting a complete axiology that satisfies both Weak Pareto and Finite Anonymity is to invoke an *ultrafilter*, a particular kind of non-constructive object that in the present context would tell us which infinite subsets of the set of possible locations should be treated as 'large' and which as 'small'. (The resulting principle will then always prefer, for instance, to provide a given benefit to a 'large' set of locations rather than a 'small' set.) But it is very hard to imagine what could single out any particular ultrafilter, among the uncountably many that can be imposed on the infinite set of possible locations, to play this privileged axiological role.

[33] This constitutes a rejection of both Unrestricted and Finite Anonymity, and so clearly abandons risk-neutral totalism's commitment to impartiality. This sort of partiality toward nearer locations has been defended as necessary for the evaluation of infinite futures—see for instance Koopmans (1960). But note that *time* discounting alone does not avoid the problems associated with a *spatially* infinite universe; so to avoid all of the difficulties of the infinite setting, one might need a spatial as well as a temporal discount rate. For a survey of arguments against pure time discounting, see Greaves (2017a: sec. 7).

irrelevant to what we should do and plays no essential role in guiding our practical decisions (cf. Taurek 1977). Perhaps the value of outcomes is agent-relative, incorporating strong partiality toward the agent and their nearest and dearest with little if any weight given to far-off strangers, or depending entirely on the agent's subjective preferences. Or perhaps outcomes don't even have agent-relative value, and which of your options you should prefer in a given choice situation is determined by thoroughly non-axiological considerations.

3. **Narrow person-affecting views**: The overall value of an outcome, from the perspective of an agent in a particular choice situation, depends only on those locations that exist *necessarily* with respect to that choice situation, i.e., regardless of the agent's choice (see, e.g., Temkin 1987: 166–167).[34]
4. **Ignoring small probabilities**: Sufficiently low-probability states or outcomes should simply be ignored in ranking prospects; prospects should be valued at their expected total value, conditional on such low-probability states or outcomes not occurring.[35] Suitably formulated (which is no small challenge—see Kosonen n.d.), this policy of small probability neglect might allow us to ignore the speculative hypotheses (like cosmological natural selection and future supertask computers) that allow our actions to predictably affect infinitely many locations, and thereby rescue us from widespread incomparability in real-world decision situations.

Compared to risk-neutral totalism, on any of these views, the case for longtermism appears weaker. On the other hand, each of these views has serious drawbacks—in our view, greater than those of the various proposed extensions of risk-neutral totalism in the infinite setting. But no doubt some will disagree, and it is undeniable that the challenges of infinite axiology do count somewhat in favour of normative worldviews less favourable to longtermism.

7 Conclusion

We set out to investigate whether the axiological challenges of infinite worlds undermine the risk-neutral-totalist case for longtermism. The results of this investigation are, unfortunately, mixed and uncertain.

Our own provisional conclusions are as follows. First, any plausible extension of risk-neutral totalism to infinite contexts can rank prospects in any decision where our choices affect only finitely many local prospects. In such decisions, any such view preserves the risk-neutral-totalist case for longtermism by letting us ignore all those locations whose prospects are unaffected. And many of our real-world decisions—particularly those involving no predictable long-term effects—will plausibly have this nice character.

[34] This is a species of agent-relative consequentialism, but an especially notable one for present purposes. A very similar view could be articulated in *time-relative* rather than *agent-relative* fashion: the overall value of an outcome, from the perspective of a particular moment in time, depends only on those locations that exist at that time, or on those locations whose existence is nomologically necessary given the state of the universe at that time. Such time-relative views violate both Unrestricted and Finite Anonymity, and have some deeply counter-intuitive implications (see Greaves 2017b: 8–9).

[35] This view is dubbed *Nicolausian discounting* by Monton (2019), who defends it. For objections, see for instance Wilkinson (2022) and Beckstead and Thomas (2024).

Second, we should assign some non-zero probability to physical hypotheses that let us predictably affect infinitely many locations in certain decisions. This means that our choices—at least those that affect the long-run future—can have at least some small effect on infinitely many local prospects.

Third, in those circumstances, any otherwise plausible extension of risk-neutral totalism that makes comparisons (rather than implying widespread incomparability) will very likely preserve the risk-neutral-totalist case for longtermism. Indeed, it seems that it would even strengthen that case by implying that the long-term stakes of our actions are infinite.

Fourth, if otherwise plausible extensions of risk-neutral totalism instead imply widespread incomparability in practice, then we plausibly have good reason to reject risk-neutral totalism. And various impossibility results in infinite axiology might also be taken to motivate the rejection of risk-neutral totalism, since they imply that at least some of its attractive features in finite contexts must be given up in infinite contexts.

We ourselves are inclined to think that risk-neutral totalism remains more plausible than each of the alternatives raised above, despite the impossibility results.[36] And we hold out hope that the correct extension of risk-neutral totalism to infinite contexts, while it may countenance some incomparability, will not imply very widespread incomparability in real-world choice situations. But this hope has not yet been fully vindicated—it is not yet clear what the correct extension is. (Nor has it been vindicated, nor the correct extension identified, for the many axiologies other than risk-neutral totalism that are also favourable to longtermism.) Until that correct extension is found, while infinitary worries about the case for longtermism can be mitigated, they cannot be totally allayed.[37,38]

References

Arntzenius, F. (2014), 'Utilitarianism, Decision Theory and Eternity', in *Philosophical Perspectives* 28/1: 31–58.

Askell, A. (2019), *Pareto Principles in Infinite Ethics*, PhD thesis, New York University.

Bader, R. (n.d.), *Person-Affecting Population Ethics* (unpublished manuscript).

Basu, K. and Mitra, T. (2007), 'Utilitarianism for Infinite Utility Streams: A New Welfare Criterion and Its Axiomatic Characterization', in *Journal of Economic Theory* 133/1: 350–373.

Beckstead, N. and Thomas, T. (2024), 'A Paradox for Tiny Probabilities and Enormous Values', in *Noûs* 58/2: 431–455.

Bernoulli, D. (1738/1954), 'Exposition of a New Theory on the Measurement of Risk', in *Econometrica: Journal of the Econometric Society* 22/1: 23–36.

[36] We both incline at least somewhat toward totalism. One of us (HW) also inclines toward risk neutrality, while the other (CT) does not, but thinks that the correct principles for evaluation of risky prospects will have similar implications in practice (see Tarsney, forthcoming).

[37] For interested readers, we have two suggestions for future research. First, compared to the extensive literature on the evaluation of infinite outcomes, there has been relatively little work in infinite axiology on the evaluation of prospects. More such work, exploring possible strengthenings of principles like SVPD and their practical implications, could be very useful. Second, most views in infinite axiology make use of an identity or counterpart relation across possible outcomes, and the practical implications of these views depend on the nature of that relation. But most work in this area does not incorporate a full theory of the relevant relation or think through what it implies about our real-world circumstances. This sort of work also seems essential to fully understanding the practical implications of an infinite axiology (see fn. 25 above).

[38] For helpful feedback on earlier versions of this chapter, we thank Teru Thomas, Riley Harris, Elliott Thornley, H. Orri Stefánsson, and participants at both the 2nd Oxford Workshop on Global Priorities and the 2019 Workshop on the Economics of Catastrophe in Oxford.

Bostrom, N. (2011), 'Infinite Ethics', in *Analysis and Metaphysics* 10: 9–59.
Buchak, L. (2023), 'How Should Risk and Ambiguity Affect Our Charitable Giving?', in *Utilitas*: 35/3: 175–197.
Cain, J. (1995), 'Infinite Utility', in *Australasian Journal of Philosophy*: 73/3: 401–404.
Carroll, S. M. (2020), 'Why Boltzmann Brains Are Bad', in S. Dasgupta, R. Dotan, and B. Weslake (eds.), *Current Controversies in Philosophy of Science* (Taylor & Francis).
Clark, M. (n.d.), *Infinite Ethics, Intrinsic Value, and the Pareto Principle* (unpublished manuscript).
Earman, J. and Norton, J. D. (1993), 'Forever Is a Day: Supertasks in Pitowsky and Malament-Hogarth Spacetimes', in *Philosophy of Science* 60/1: 22–42.
Greaves, H. (2017a), 'Discounting for Public Policy: A Survey', in *Economics & Philosophy* 33/3: 391–439.
Greaves, H. (2017b), 'Population Axiology', in *Philosophy Compass* 12/11: e12442.
Greaves, H. and MacAskill, W. (this volume), 'The Case for Strong Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Greaves, H. and Tarsney, C. (this volume), 'Minimal and Expansive Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Hamkins, J. D. and Montero, B. (2000), 'With Infinite Utility, More Needn't Be Better', in *Australasian Journal of Philosophy* 78/02: 231–240.
Harsanyi, J. C. (1955), 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', in *Journal of Political Economy* 63/4: 309–321.
Hong, F. and Russell, J. S. (forthcoming), 'Paradoxes of Infinite Aggregation', in *Noûs*.
Jonsson, A. and Voorneveld, M. (2018), 'The Limit of Discounted Utilitarianism' in *Theoretical Economics* 13/1: 19–37.
Knobe, J., Olum, K. D., and Vilenkin, A. (2006), 'Philosophical Implications of Inflationary Cosmology', in *The British Journal for the Philosophy of Science* 57/1: 47–67.
Koopmans, T. C. (1960), 'Stationary Ordinal Utility and Impatience', in *Econometrica: Journal of the Econometric Society* 28/2: 287–309.
Kosonen, P. (n.d.), *How To Discount Small Probabilities* (unpublished manuscript).
Lauwers, L. (2010), 'Ordering Infinite Utility Streams Comes at the Cost of a Non-Ramsey Set', in *Journal of Mathematical Economics* 46/1: 32–37.
Lauwers, L. and Vallentyne, P. (2004), 'Infinite Utilitarianism: More Is Always Better', in *Economics and Philosophy* 20/2: 307–330.
Liedekerke, L. V. (1995), 'Should Utilitarians Be Cautious about an Infinite Future?', in *Australasian Journal of Philosophy* 73/3: 405–407.
Liedekerke, L. V. and Lauwers, L. (1997), 'Sacrificing the Patrol: Utilitarianism, Future Generations and Infinity', in *Economics and Philosophy* 13/2: 159–174.
MacAskill, W. (2016), 'Smokers, Psychos, and Decision-Theoretic Uncertainty', in *Journal of Philosophy* 113/9: 425–445.
MacAskill, W., Vallinder, A., Shulman, C., Österheld, C., and Treutlein, J. (2021), 'The Evidentialist's Wager', in *Journal of Philosophy* 118/6: 320–342.
Meacham, C. (2020), 'Too Much of a Good Thing: Decision-Making in Cases with Infinitely Many Utility Contributions', in *Synthese* 198/8: 7309–7349.
Monton, B. (2019), 'How to Avoid Maximizing Expected Utility', in *Philosophers' Imprint* 19/18: 1–25.
Nover, N. and Hájek, A. (2004), 'Vexing Expectations', in *Mind* 113/450: 237–249.
Pettigrew, R. (2022), 'Effective Altruism, Risk, and Human Extinction', GPI Working Paper No. 2-2022 (Global Priorities Institute, Oxford University).
Smith, Q. (2003), 'Moral Realism and Infinite Spacetime Imply Moral Nihilism' in H. Dyke (ed.), *Time and Ethics: Essays at the Intersection* (Springer), 43–54.
Smolin, L. (1992), 'Did the Universe Evolve?', in *Classical and Quantum Gravity* 9/1: 173.
Tarsney, C. (forthcoming), 'Expected Value, to a Point: Moral Decision-Making under Background Uncertainty', in *Noûs*.
Tarsney, C. and Thomas, T. (2024), 'Non-additive Axiologies in Large Worlds', *Ergo: An Open Access Journal of Philosophy* 11/9: 238–287.
Taurek, J. M. (1977), 'Should the Numbers Count?', in *Philosophy & Public Affairs* 6/4: 293–316.
Temkin, L. S. (1987), 'Intransitivity and the Mere Addition Paradox', in *Philosophy & Public Affairs* 16/2: 138–187.
Thomas, T. (2023), 'The Asymmetry, Uncertainty, and the Long Term', in *Philosophy and Phenomenological Research* 107/2: 470–500.
Tipler, F. J. (1994), *The Physics of Immortality: Modern Cosmology, God, and the Resurrection of the Dead* (Anchor Books).

Vallentyne, P. (1993), 'Utilitarianism and Infinite Utility', in *Australasian Journal of Philosophy* 71/2: 212–217.
Vallentyne, P. and Kagan, S. (1997), 'Infinite Value and Finitely Additive Value Theory', in *The Journal of Philosophy* 94/1: 5–26.
Wilkinson, H. (2021a), *Infinite Aggregation*, PhD thesis, Australian National University.
Wilkinson, H. (2021b), 'Infinite Aggregation: Expanded Addition', in *Philosophical Studies* 178/6: 1917–1949.
Wilkinson, H. (2022), 'In Defence of Fanaticism', in *Ethics* 132: 445–477.
Wilkinson, H. (2023), 'Infinite Aggregation and Risk', in *Australasian Journal of Philosophy* 101/2: 340–359.
Wilkinson, H. (n.d.a), *Chaos, Add Infinitum* (unpublished manuscript).
Wilkinson, H. (n.d.b), *Can the Long Term Matter if Extinction Does Not* (unpublished manuscript).
Zame, W. R. (2007), 'Can Intergenerational Equity Be Operationalized?', in *Theoretical Economics* 2/2: 187–202.

7

Longtermism and the Complaints of Future People

Emma J. Curran

Introduction

We are not particularly good at reliably controlling the future. Perhaps, when looking at a horizon of a few decades, or hundred years, we can make reasonable estimates about the effects of our actions. Yet, when considering the very far future—a future stretching hundreds of thousands of years ahead of us, if not more—there are deep epistemological challenges preventing us from having confidence about the impact of our present-day actions.[1]

Despite the uncertainty surrounding our ability to shape the very far future, longtermists believe we have good moral reason to invest resources into attempting to positively shape it. The argument for longtermism is quite simple: even if anything we can do has only a very tiny chance of improving the far future, given the massive number of people who will likely exist throughout the future, in expectation, the amount of good or value we will bring about will be massive.[2] In fact, long-term interventions likely bring about many orders of magnitude more expected good than available short-term options. As such, according to Hilary Greaves and Will MacAskill, in many decision situations, the option which most improves the prospects of those living in the very far future will also be the option which brings about the most *good*, in expectation (Greaves and MacAskill 2021: 3–4).[3] Following Greaves and MacAskill, we will call this claim, henceforth, 'axiological longtermism'.

Now, if you're the sort of agent who simply cares about bringing about the most good, then insofar as the argument sketched above is sound, you have decisive reason to divert your resources into improving the far future. Yet, many people don't simply care about bringing about the most good or value. Many people care about discharging their moral obligations. As such, many people don't simply want to know if axiological longtermism is true, they want to know if this gives rise to a moral obligation to invest in long-term interventions. They care about the truth of 'deontic longtermism', the claim—very roughly

[1] See Tarsney (2023) for discussion.

[2] Many think there is something problematic about allowing tiny probabilities of giant payoffs to dictate our actions. See Bostrom (2009), Balfour (2021), Wilkinson (2022), and Kosonen (2023) for a discussion of the plausibility of 'fanatical' decision theories which give rise to this results.

[3] See Thorstad (forthcoming) for a discussion of the scope of cases, or decision situations, in which this claim might hold true.

Emma J. Curran, *Longtermism and the Complaints of Future People* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Emma J. Curran 2025. DOI: 10.1093/9780191979972.003.0007

put—that in many decision instances, we have a moral obligation to choose the option which most improves the prospects of those living in the far future.

In this chapter, I aim to explore contractualist assessments of the veracity of deontic longtermism. However, before doing so, I must make some qualifications. When it comes to discussions of the claims of future persons, an initial, and seemingly unavoidable, hurdle is the non-identity problem. The non-identity problem, put simply, is that future people seemingly cannot have claims to our assistance grounded in their well-being insofar that such assistance would change the identity of future populations and they, without our assistance, would have lives worth living (Parfit 1982; 1984: ch. 16). A discussion of the non-identity problem and contractualism, however, is absent from this chapter, in large part because I think such a discussion is unnecessary. In this chapter, I will assume that future people generate the same sort of complaints regarding their well-being as presently existing people do (Kumar 2003a; Hare 2007). I will show that, even granting this assumption, we have good reason to be suspicious of justifying long-term intervention on the basis of future people's claims to them.

A second qualification is that throughout this chapter, I will assume, when assessing long-term interventions, that the population size remains fixed across the future. Most interventions which shape the far future will have an impact, in some way, on the size of the number of people in the future. However, to take variable population sizes into account when assessing such interventions is to enter debates about the value of bringing people into existence, which would take me far beyond the scope that this chapter permits. A complete contractualist account of the ethics of longtermism will require an account of variable population sizes. However, for the sake of simplicity, I will abstain from such a discussion.

I will proceed, in section 1, to highlight an argument which I have developed elsewhere in a slightly different form (Curran 2025). The argument is to the effect that contractualism, and complaints-based theorising in general, assesses many long-term interventions as generating extremely weak claims in their favour. As such, rather than being obligatory, investing in long-term interventions will, in many cases, be impermissible on a contractualist assessment. In section 2, I will consider a number of objections to this argument, and finally, in section 3, I will briefly outline the prioritisation recommendation that contractualism might have.

1 An argument against deontic longtermism

Scanlonian contractualism tells us that an action is right iff it is permitted by a principle which no one could reasonably reject. In unpacking what it means for an individual to 'reasonably reject', contractualism makes use of two types of restrictions. The first is the 'personal-reasons restriction' (Scanlon 1998: 218–223; Kumar 2003b). Individuals can only legitimately reject principles on the basis of the principle's impact on them, as opposed to on the basis of 'impersonal' concerns such as its impact on the sum, or distribution, of well-being in the world. For simplicity's sake, I will say that, in a decision situation, an individual has a complaint if her well-being would be higher under one of the available options.[4] The

[4] This definition of 'complaint' implicitly accepts the 'personal reasons restriction' of Scanlonian contractualism.

strength of an individual's complaint against a policy is a function of the impact the policy would have upon her well-being.

The second restriction contractualism contains is the 'individualist restriction'. In decision situations in which all options have complaints against them, agents cannot consider the aggregate complaint against any option. Rather, each complaint must be considered individually, and go through a process of pairwise comparison against the complaints with which it competes. Complaints 'compete' against each other insofar as they are mutually unsatisfiable.

Presented with two options, the first to save 100 people from a headache, and the second to save one other person from death, contractualism tells us that the agent is not permitted to consider the weight of all 100 complaints of a headache together. Rather, each complaint of a headache must be compared individually against the complaint of death. The individualist restriction is designed precisely to avoid counter-intuitive verdicts in which we are obligated to save a large number of people from a trivial harm—like headaches—instead of a small number of people from a significant harm—like death.[5]

How might contractualism assess deontic longtermism? I ask you to consider the following:

> **Chancy Treatment:** There are 100,000,000 people at a 0.005% risk of developing a fatal disease. We can treat the group with an inexpensive pill which will reduce their risk of the disease to 0%. Unfortunately, Arthur has also contracted the fatal disease. However, we can have 99.999% confidence of curing him through an expensive treatment. Due to our finite resources, we can only (a) treat the entire group, or (b) treat Arthur, not both.

It should be fairly clear that (a) is an analogue for long-term interventions; it brings about a lot of good in expectation—indeed, saving 5,000 lives—but does so by improving the prospects of a large group by a very tiny amount. (b) looks like, at least *some* of, the short-term interventions available to us.

What does contractualism tell us to do in this case? This depends on the account of how to assess complaints under risk that it chooses. There are two options discussed in the literature: the *ex ante* interpretation and the *ex post* interpretation. One classic formulation of the difference between these two perspectives is with regards to the good under distribution; the *ex ante* perspective is concerned with the distribution of individual prospects, whilst the *ex post* perspective is concerned with the pattern of *ex post* outcomes.

An individual's *ex ante* complaint against a policy is simply the difference between their expected well-being given the policy was implemented and their expected well-being given the policy wasn't implemented. *Ex post* complaints, on the other hand, focus on the likelihood that someone—not that any particular person—will incur a harm or receive a benefit. The number of *ex post* complaints against a policy will be the number of people who could be harmed by the policy, with each complaint discounted by the difference in the improbability that the policy made to their incurring this harm. To illustrate, consider:

> **Side Effects:** There are 1,000,000 people at risk of developing a fatal disease, and we know that, if untreated, exactly one person will die—though it is impossible to determine who.

[5] A classic example of this type that Scanlon used to motivate the individualist restriction was the 'Transmitter Room' case (1998: 235).

> We can treat the group with a pill which will eliminate each member's risk of death. The pill causes nausea. The cost of nausea is judged to be greater than the benefit of having one's risk of the disease reduced from 0.001% to 0%, but the cost of actually having the disease is far greater than the cost of nausea.

What are the claims *ex ante*? Well, the beneficiaries in question are each individual member of the group of 1,000,000, each of which has an interest in not having the pill; their expected well-being given they take the pill is lower than their expected well-being given they do not. So, *ex ante*, we have 1,000,000 claims against the policy being implemented, each of which is proportional to how much worse off they expect to be made by having the pill.

Ex post, we can compare the complaints of the patients given the intervention is administered and given it is not administered. If the intervention is not administered, it generates one complaint of certain death. If the intervention is implemented, it generates 1,000,000 complaints of nausea.

The example of Side Effects should also make clear that the distinction between *ex ante* and *ex post* is not an ethically unimportant one. Indeed, in Side Effects, the *ex ante* and *ex post* perspectives both capture important and seemingly contradictory considerations. In one way of looking at the world, the intervention is in no one's interests, but from another perspective the intervention is of fundamental importance to someone's interests—namely the one who would die if not treated.[6]

To further clarify the evaluation of *ex post* complaints, consider a variant of Side Effects in which, in expectation, one person will die if not treated, but the risk is independent. In this case, as potentially all 1,000,000 individuals could contract the disease, there are 1,000,000 complaints of death against not administering the intervention. However, each complaint is discounted at a different rate, following a binomial distribution.

Now, what would *ex ante* contractualism say about Chancy Treatment? Well, the complaint against (a) —the long-term analogue—is one complaint of, near certain, death. On the other hand, (b) generates 100,000,000 complaints of a 0.005% risk of death. Going through a process of pairwise comparison, clearly the complaint of near death will outweigh each of the 100,000,000 complaints of a tiny risk of death.

The most important upshot is that, given the fact that long-term interventions can only improve the prospects of each future person by a very tiny amount, on an *ex ante* account, they will generate extremely weak individual claims in their favour, indeed, much weaker than the claims generated by some available short-term interventions. Leaving aside the issue of whether there is a genuine competition between long-term and short-term interventions, this demonstrates that—at least on an *ex ante* contractualist account—if we were to engage in long-term intervention, it would be because of the strong claim generated by a correlative short-term intervention. Not, *pace* the longtermist, because of the value brought about by the long-term intervention.

[6] As such, the intervention in Side Effects is an example of what Stephen John calls an 'absolute prevention paradox': an intervention which is, in the *ex ante* sense, in no one's interest, yet in the *ex post* sense, in the interests of some (2014: 30–34).

Next, let us consider an *ex post* contractualist assessment of Chancy Treatment. It is tempting to think interventions like (a) in Chancy Treatment are easily justified *ex post* by appealing to those people who we expect to bestow a good upon by engaging in the intervention, whomever they may be. Consider again Chancy Treatment; we can offer an *ex post* justification for engaging in the preventative option by appealing to the complaints of the 5,000 individuals we expect to save from death (or more precisely, the 100,000,000 complaints of death discounted following a binomial distribution, equivalent to roughly 5,000 complaints of death.) Contractualism permits, if not obligates, us to save the greatest number—even if it doesn't allow aggregating complaints (Kamm 1993: 101–119; Scanlon 1998: 232–233). As such, one might think contractualism would permit, if not obligate, us to engage in the preventative intervention, (a), in Chancy Treatment.

This line of thought, however, is mistaken. Option (a) in Chancy Treatment features an important disanalogy from many important long-term interventions. Unlike Chancy Treatment, a great many of the risks longtermists are trying to mitigate are not independent. This is especially true of interventions which seek to mitigate catastrophic risks, extinction risks, or suffering risks. The risks that longtermists are concerned with of unaligned artificial intelligence (AI), climate change, supervolcano eruptions, or asteroid impacts are not independent. Indeed, each individual's risk of being harmed by each of these catastrophes is deeply connected. Either the supervolcano erupts, thereby killing however many people, or it doesn't erupt, meaning no one is killed by it. Likewise, either an asteroid hits or it doesn't, and AI goes rogue, or it does not.

The dependent nature of the risks facing each individual in catastrophe scenarios fundamentally changes the nature of the complaints that catastrophic risk interventions generate in their favour. Imagine a treatment of Chancy Treatment—Chancy Treatment (Dependent)—in which the 0.005% risk isn't independent but rather dependent in the same way as the catastrophic risks discussed. In this case, it wouldn't be true to say we reasonably expected to save 5,000 lives by engaging in option (a). This is because there is no state of affairs in which 5,000 people die from the disease. Either no one dies, or all 100,000,000 die. Intervention (a) simply reduced the likelihood of such an eventuality. As such, *ex post*, we don't have 5,000 complaints of death. We have 100,000,000 complaints of death, but each is discounted by the difference in the improbability that the intervention made to the harm actualising. That is, it is discounted by 99.995%.

It should be clear that once we properly conceptualise the nature of the claims generated by many long-term interventions, even from the *ex post* perspective they are extremely weak. Certainly not strong enough to compete with available short-term interventions, and certainly not strong enough to 'dictate' how we ought to distribute our resources.

So, it seems that so long as contractualists maintain the personal reasons restriction, on both the *ex ante* and *ex post* perspective, long-term intervention (in many cases) will generate extremely weak individual claims in their favour. And, if the contractualist also maintains their individualist restriction, these weak individual claims will unlikely be decisive when considering the various options available to us to help others.[7] This is generalisable to all complaints-based moral theories; insofar as they limit claims to personal welfare

[7] In fact, in other work I have demonstrated that even if complaint-based moral theorists weaken the individualist restriction—for example to permit limited or partial aggregation (Voorhoeve, 2014; 2017)—then, on both an *ex ante* and an *ex post* perspective, long-term interventions will still look extremely uncompetitive.

reasons, and also do not permit the aggregation of such claims, then the claims that long-term interventions will generate will be extremely weak.

2 A *reductio* for contractualism

For the committed contractualist, the argument sketched in section 1 might be decisive reason to give up their longtermist activities. However, I'm sure to many readers the argument has the distinct flavour of a *reductio*. This is especially true when we consider the further implications of these conclusions. *Ex ante,* the long-term analogue in Chancy Treatment looks a great deal like many population-level healthcare policies we engage in currently. Moreover, on a contractualist account, regardless of whether you adopt an *ex ante* or an *ex post* perspective, interventions analogous to Chancy Treatment (Dependent) seem to be ruled out. This, in particular, might be too counter-intuitive to countenance. Consider:

> **Natural Disaster:** Above a city of 10,000,000 people there is a volcano which is at a 0.05% risk of exploding and killing the population of the city. We can reduce the risk of the population of the city dying, in the case of the volcano's eruption, to 0%, through an expensive engineering project. Elsewhere, in a faraway city, there is a patient, Donald, who has contracted an illness which will prove fatal unless we choose to give him an expensive treatment. Due to our finite resources, we can only treat Donald or tend to the volcano risk, not both.

It seems sensible for a government to tend to such risks of natural disaster. In fact, it seems that governments are morally obligated to attend to such risks even if doing so means diverting some of our finite resources away from aiding individuals who are at high risk of a serious harm. However, the structure of Natural Disaster mirrors that of Chancy Treatment (Dependent), and, as we have seen, on both an *ex ante* and an *ex post* perspective, contractualism compels us to choose the option to help Donald. It is tempting to think that any minimally plausible moral theory which informs our priority setting ought to allow us to tend to natural disaster risk. As such, contractualism—and other similarly anti-aggregative complaint-based moral theories—fails to be even minimally plausible.

I agree that this is a hard bullet for the contractualist to bite, and it certainly does not bode well for the practicability of contractualism if this is a genuine implication of the theory. To avoid such a conclusion, the contractualist needs to point to a disanalogy between Natural Disaster and long-term interventions which will explain why we might be able to attend to natural disaster risks. I claim that such a disanalogy lies in the fact that we can reasonably expect to save a life (or prevent some harm) through mitigating natural disaster risks in a manner we can't when it comes to catastrophic risk. This opens the door to a potential *ex post* justification of natural disaster mitigation schemes.

The first justification I want to offer operates on the level of 'programmes' of interventions. That is, we can justify natural disaster interventions by evaluating larger decisions, such as general policy decisions, rather than specific interventions. Let's—for the moment—grant that each particular natural disaster intervention might look a lot like the

preventative intervention in Chancy Treatment (Dependent). As such, we will assume that we do not reasonably expect *any given* intervention to save any lives. Nonetheless, plausibly, when we engage in a programme of such interventions we really do expect to save lives at some point. This is, in part, due to the fact that natural disasters have a high antecedent chance of actually occurring at some point and that programmes of mitigation can decisively lower the probability of fatal natural disaster events such that we expect to actually prevent the deaths of some people. This allows us to, *ex post*, justify a programme of natural disaster risk mitigation by way of the significant claims of those people whose lives we expect to save, even if it might not be reasonable for us to expect each particular intervention in the programme to save a life.[8]

For illustrative purposes, this justification paints decisions to engage in natural disaster mitigation intervention as looking like:

> **Lottery:** You can save one life or receive a single ticket for a lottery of N-tickets. If your ticket is drawn, 100,000 people, who otherwise would have died, will be saved.

Given an appropriately specified N[9], whilst the expected value of engaging in the lottery is larger than saving the single life, you don't reasonably expect that engaging in the lottery will have any impact on the outcome of the 100,000 people. So, *ex post* contractualists will favour saving the one over attempting to save the 100,000 as each of their complaints of death will be discounted significantly. However, the decision to engage in a programme of natural disaster management would be analogous to:

> **Grand Lottery:** You can save X lives or receive one ticket for each of X lotteries of N-tickets. If your ticket for a lottery is drawn, then 100,000 people, who otherwise would have died, will be saved.[10]

In this case, for some X and N, you can reasonably expect that by engaging in the grand lottery scheme, you will save the 100,000 people who otherwise would die. As X gets closer to N, the probability that you will save all 100,000 people will increase, meaning the *ex post* complaints of death in favour of saving the 100,000 will be discounted to a lesser degree. Eventually these *ex post* complaints will be discounted to a small enough extent that they will be sufficiently near certain to allow them not to be outweighed individually by the one complaint of death. If this analysis is correct, then it saves the *ex post* theorist from troubling incorrect verdicts in cases like Natural Disaster.

[8] This explanation turns around the fact that *ex post* justifications do not decompose; on the *ex post* perspective, the deontic status of a group of actions is different from the status of each individual sub-action. Whilst this is usually held as a troubling implication of *ex post* theories (cf. Hare 2016), it allows the *ex post* theorist to avoid counter-intuitive verdicts in cases like Natural Disaster.

[9] This will involve it being strictly smaller than 100,000 but larger than, for now let's arbitrarily say, 3. Each of the *ex post* complaints of death will be discounted by (N-1)/N.

[10] Note, I have designed this case to look at programmes of natural disaster mitigation which look at different risks which face different people, for example, a programme to mitigate the harm of flooding in multiple locations around a country. However, a similar line of justification can be offered to justify a programme of interventions which all seek to mitigate the same risk, for example, a programme of interventions all targeting the flood risk in a specific town in a country. In this case, rather than receiving one ticket for X different lotteries, the agent would receive X tickets for one lottery.

At this point, one might wonder if by paving the way for interventions like the one in Natural Disaster to gain justification on *ex post grounds*, this line of thought might also justify a programme of catastrophic risk mitigation schemes. The thought here is that, if you group together a sufficient number of long-term interventions, which look like the one in Chancy Treatment (Dependent), then at a certain point you will have good statistical reason to believe that one of them will be successful. As such, you will also have good statistical reason to believe that you will save a great number of lives, generating significant *ex post* complaints in support of this programme of interventions.[11] However, I would point out that this line of justification is likely impractical for typical catastrophic risk interventions. Given the improvements to probabilities associated with long-term interventions are so tiny, the number of interventions such a programme would need to include to generate near statistical certainty of savings lives would be, at a minimum, practically preclusive. As such, at least at first glance, we seem to have a line of reasoning which allows the contractualist to avoid the counter-intuitive verdicts in cases like Natural Disaster, which, nonetheless, does not commit the contractualist to permitting investment in catastrophic risk interventions.

Whilst this line of thought is promising, it does face a series of difficult problems. The first is that, at least on the contractualist account we have been considering, the application of this justification is limited simply to cases in which the probability of not saving people becomes very small. Whilst we might be able to treat a very slightly discounted complaint of a harm as equivalent in strength to undiscounted complaints of the same harm, there is clearly a limit to how discounted a complaint can be for us to do this. Yet, there is intuitively a gap between interventions in which we do not reasonably expect to save any lives and interventions in which we are *almost* certain to save a life. Consider, for example, a programme of natural disaster risk mitigation which makes it the case that we have 70% confidence of saving lives. This seems like the sort of programme of interventions which should not be disqualified out of hand. But a complaint of death discounted by 30% is just not equivalent to undiscounted complaint of death. So, this might be a weaker response to the problem than we had hoped; it is unable to justify a large number of intuitively permissible interventions.

I will quickly gesture at two potential responses to this problem. The first is to simply accept the limitations of this solution, whilst highlighting that at least it allows the contractualist to justify those classes of interventions which would be most counter-intuitive to exclude. This at least lessens the bite of the initial criticism. The second response is to highlight that the power of this justification can be extended significantly if we were to weaken one of the contractualist restrictions.

A number of complaints-based theorists have weakened the individualist restriction such that the aggregation of some complaints would be permitted. These 'partially aggregative' theories allow aggregation in cases in which the lesser competing complaint is, nonetheless, of a similar moral significance to the complaint against which it competes (see Voorhoeve

[11] Whilst the likelihood of any catastrophic risk intervention being successful is unlikely to be totally independent—for example, multiple interventions are likely to share similar assumptions and be based on similar models—it is also not the case that the success of catastrophic risk interventions is all or nothing either, such that all interventions work or none do (unlike the outcomes of a catastrophic risk itself, which, as discussed in section 1, do have this all or nothing structure).

2017). Whilst a discussion of such accounts is beyond this chapter, the contractualist could certainly use such a stance to justify *ex post*, for example, the intervention which would give a 70% chance of saving lives from a natural disaster.[12] In this case, the *ex post* complaints of death, discounted by 30%, would be allowed to aggregate together to compete against the lives competing interventions could certainly save. Provided the state of affairs which the intervention prevented with 70% likelihood contained enough deaths, the intervention will have a large aggregate complaint in its favour.

Moving onto a second problem for the programme-level justification, it seems that such a move would allow us to justify interventions by tacking them onto already *ex post* justifiable programmes. Take a set of interventions, S, which we can justify *ex post* on the basis of the high likelihood that together they will save lives from natural disasters. Now consider an additional individual intervention, let's call it i. So long as this intervention: (i) does not increase the likelihood of people dying from natural disasters such that we no longer expect to save lives through the programme; and (ii) does not come with an opportunity cost which generates an equivalent or greater number of complaints of equivalent or greater strength than is in favour of the programme of interventions; then we can add i to S to form another set, S*, and S* will also be justified *ex post*. This means that interventions, like i, can piggyback on the justifiability of programmes of interventions without contributing anything themselves.

In light of these problems, I would like to gesture at another means for the contractualist to justify the permissibility of natural disaster intervention (and the like). This justification, unlike our last, occurs on the level of individual interventions. Crucially there seems to be a disanalogy between natural disasters and long-run catastrophic risks which Chancy Treatment (Dependent) has obscured—namely, the source of the low probabilities of saving a life. As noted previously, the reason we can only improve the prospects of future people by a very tiny amount, or reduce the likelihood of a catastrophic state of affairs by a very small amount, is because we lack the power to reliably control the future. It is not to do, principally, with the antecedent likelihood that catastrophe will occur in the very far future. The likelihood might be very high—regardless, we just can't do that much about it.

On the other hand, the source of the low probabilities associated with the likelihood of natural disaster interventions actually preventing disaster is not so closely tied with our inability to mitigate present-day natural disasters. Of course, there are some natural phenomena which we might struggle to effectively prevent or mitigate. But there are many natural disasters for which we have management methods which definitely work. This is especially clear with mitigation schemes; take, for example, seismic designing, which reliably saves lives during earthquakes by minimising damage to buildings. Or consider a slightly more realistic version of Natural Disaster; we can confidently reduce mortality, in the case of a volcanic eruption, through city and evacuation planning.

In such cases, the low chance of saving the lives of individuals from natural disaster does not result from our lack of confidence about the efficacy of the intervention, but rather the antecedently low chance of the specific catastrophe occurring for the intervention

[12] For discussion of partially aggregative theories and longtermism, see Curran (2025). There I argue that partially aggregative theories run into the exact problems that fully non-aggregative theories, like contractualism, run into on both the *ex ante* and *ex post* levels.

to mitigate. The different sources of the low probabilities associated with natural disaster management and catastrophic risk mitigation are important. Namely, the probabilities associated with saving lives through catastrophic risk schemes won't grow if you extend your time frame; we don't become more confident that our action will have some large, net positive impact when we consider the whole course of history. On the other hand, this is the case for certain types of natural disaster management. Given that the probability of these interventions saving lives is currently bound by the probability of the disaster occurring, the higher the likelihood of the disaster occurring at some point in our time frame, the greater the probability that such interventions will save lives. Therefore, by extending the time frame relevant to our intervention assessment, such natural disaster management schemes become increasingly justifiable *ex post*. That is, individual natural disaster interventions can be *ex post* justifiable because we really do expect them to save lives at some point in the long run, in a way that we don't reasonably expected catastrophic risk interventions to do at some point.

This is a peculiar result, insofar as it shows the contractualism can favour some long-term interventions—at least on the *ex post* perspective. But importantly, contractualism does not favour the characteristic long-term interventions which bring about vast amounts of goodness by tending to tiny probabilities (or risks) of enormously good (or bad) outcomes. Instead, contractualism can favour those interventions which, through the perspective that longtermism brings, decisively and reliably attend to the high probability that many significant complaints of a harm will occur at some point in the future.

To briefly conclude this section, it seems that contractualism can—to some extent—avoid the bite of the analogue between presently accepted activities like attending to natural disaster risk and activities it judges as generating weak claims in its favour, such as (a) in Chancy Treatment (Dependent). Whilst this might have disarmed one of the reasons pointing to reading section 1 as an argument against contractualism itself, this is limited. Given our inability to control the far future is pervasive, contractualism seems—at least on the *ex ante* perspective—to largely preclude attending to the well-being of future people. This itself might seem like too large a loss to accept.

3 Justifiability and prioritisation

Having surveyed a range of problems facing the contractualist critique of longtermism, I would, in the final section, like to highlight some of the insights contractualism might have for prioritisation within the longtermist movement.

The prospects for long-term interventions on the *ex ante* perspective might seem dire given the discussion in section 1. As a result of the low level of confidence we can have in the efficacy of many very long-term interventions, these interventions will generate extremely weak claims in their favour. One upshot of this is that the *ex ante* contractualist who cares about helping the far future will be pushed into focusing on 'sure thing' interventions, which can bestow benefits on future people with a good level of confidence.

There are at two observations to accompany this point. First, given the epistemological issues we face when it comes to predicting our impact on the future, there are significant questions to be answered about the size of the set of interventions which can bestow interventions on the far future with any degree of confidence. One fairly straightforward

example would be to leave money in trust for the future. Another possible class of interventions which we might have higher confidence about is that of those which aim to 'bring forward progress'.

Within longtermist circles, there has been some discussion of the value of bringing forward progress. The idea is that, seemingly, the conditions in which humans live have been improving over time, and such improvements, some believe, are very likely to continue in the future. As such, if we can bring forward humanity's progress—even by a small amount—then the average well-being of each human will be higher at a given time than it would be otherwise. In comparison to other long-term interventions, it might be thought that we can have at least reasonable confidence in bestowing these benefits; there are factors which, at least at first glance, seem to contribute straightforwardly to humanity's progress, including funding into research, alongside expenditure on population-level health and literacy.[13]

This leads us to the second observation. Whilst there may be a small range of long-term interventions about which we have relatively high confidence, these interventions may still struggle to be competitive on the *ex ante* contractualist model. This will be the case if such interventions trade off the possibility of bestowing large benefits on beneficiaries in order to achieve larger chances of bestowing *any* benefit upon its beneficiaries. In this case, the *ex ante* complaints in favour of such interventions are weak, not because they are discounted heavily, but rather because the benefits they bestow are meagre—or at least seem so in comparison to the sorts of benefits we can bestow with available short-term interventions.

Plausibly, interventions which seek to bring forward the 'march of progress' might look very much like this. Assuming we can bring forward progress with a higher degree of confidence, the value of doing so might be fairly trivial for each person. Let's say we bring progress forward by five years. Whilst in aggregate this might bring about a lot of value, for each individual the improvement to their well-being will plausibly be only slight. As such, the claims each individual will have to this intervention will likely also be fairly weak.

The *ex post* contractualist enjoys more room to favour long-term interventions than her *ex ante* cousin. Nonetheless, they will only be able to engage in long-term interventions insofar as the intervention makes a significant difference to the probability of the benefit being bestowed, or the harm being incurred. This is because the *ex post* complaints are discounted by the difference in the improbability that the intervention bestowed a benefit. As such, unless the intervention significantly changes the chances that the group of future people are facing, the *ex post* complaints generated by the intervention will be steeply discounted.

One appealing upshot of the *ex post* contractualist assessment is that it, effectively, precludes agents from engaging in interventions which aim to minimise risk which is already very low. This is because, provided the antecedent risk is small enough, even if the intervention reduces the risk to zero, it will have still only changed the probability by a small amount. Thus the *ex post* claims generated by the intervention will be severely discounted. This blocks the justification of interventions—such as Pascal's Mugging (Bostrom, 2009)—which have high expected value due to their promise to mitigate a tiny risk of a catastrophically bad outcome. By a similar token, *ex post* contractualism views those interventions

[13] I use this intervention type simply as an example, placing to the side the obvious historiographical concerns regarding attempts both to perceive progress in history and to make inferences about the shape of the future.

which increase an already very high likelihood of an extremely good outcome as extremely uncompetitive.

4 Conclusions

In this chapter, I have given an account of the contractualist assessment of longtermism. I have outlined an argument demonstrating that contractualism, on both the *ex ante* and *ex post* perspectives, will view long-term interventions as uncompetitive in the context of available alternative interventions. I have then considered the view that such a conclusion speaks more to the falsity of contractualism than to that of longtermism. In particular, I have considered the claim that by precluding long-term interventions which seek to tend of catastrophic risk, contractualism arrives at a deeply counter-intuitive verdict concerning a number of, seemingly permissible, present-day behaviours. I have attempted to draw a distinction between long-term interventions targeting catastrophic risk and short-term interventions managing natural disaster risk on two fronts. First, I argued that, at least on *ex post* grounds, natural disaster management can be justified on the level of programmes of interventions, as opposed to individual interventions. Whilst this justification is open to present-day natural disaster management, it does not seem available to long-term interventions. Following a survey of potential issues for this account, I outlined a second distinction between the two interventions concerning the source of their low probabilities associated with helping any given person. Having now defended, to some extent, contractualism from this chapter's *reductio*, in section 3 I consider what implications contractualism might have for prioritisation within the longtermism movement.

References

Balfour, D. (2021), 'Pascal's Mugger Strikes Again', in *Utilitas* 33/1: 118–124.
Bostrom, N. (2009), 'Pascal's Mugging', in *Analysis* 69/3: 443–445.
Curran, E. (2025). 'Longtermism and Aggregation' in Philosophy and Phenomenological Research 110/3:1137–1151.
Greaves, H., and MacAskill, W. (2021), 'The Case for Strong Longtermism', GPI Working Paper No. 5-2021 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf, accessed 10 November 2022.
Hare, C. (2007), 'Voices from Another World: Must We Respect the Interests of People Who Do Not, and Will Never, Exist?', in *Ethics* 117: 498–523.
Hare, C. (2016), 'Should We Wish Well to All?', in *The Philosophical Review* 124/4: 451–472.
John, S. (2014), 'Risk, Contractualism and Rose's "Prevention Paradox" ', in *Social Theory and Practice* 40/1: 28–50.
Kamm, F. M. (1993), *Morality, Mortality*, vol. 1 (Oxford University Press).
Kosonen, P. (2023), 'Tiny Probabilities and the Value of the Far Future', https://petrakosonen.files.wordpress.com/2022/07/chapter-6-tiny-probabilities-and-the-value-of-the-far-future.pdf, accessed: 29 January 2023.
Kumar, R. (2003a), 'Reasonable Reasons in Contractualist Moral Argument', in *Ethics* 114/1: 6–37.
Kumar, R. (2003b), 'Who Can Be Wronged?', in *Philosophy & Public Affairs* 31/2: 99–118.
Parfit, D., (1982), 'Future Generations: Further Problems', in *Philosophy & Public Affairs* 11/2: 113–172.
Parfit, D., (1984), *Reasons and Persons* (Oxford University Press).
Scanlon, T. M. (1998), *What We Owe to Each Other* (Belknap Press of Harvard University Press).
Tarsney, C. (2023), 'The Epistemic Challenge to Longtermism', in Synthese 201/6: 1–37.
Thorstad, D. (forthcoming), 'The Scope of Longtermism', in Australasian Journal of Philosophy.

Voorhoeve, A. (2014), 'How Should We Aggregate Competing Claims?', in *Ethics* 125/1: 64–87.
Voorhoeve, A. (2017), 'Why One Should Count Only Claims with Which One Can Sympathize', in *Public Health Ethics* 10/2: 148–156.
Wilkinson, H. (2022), 'In Defence of Fanaticism', in *Ethics* 132/2: 445–477.

8

Against a Moral Duty to Make the Future Go Best

Charlotte Franziska Unruh

1 Introduction

> Humanity, today, is like an imprudent teenager. The most important feature of the most important decisions that a teenager makes, like what subject to study at university and how diligently to study, is not the enjoyment they will get in the short term, but how those decisions will affect the rest of their life. (Greaves and MacAskill 2025: 17)

This analogy given by Greaves and MacAskill suggests that we, the present generation, are sacrificing our long-term potential to short-term pleasure. In doing so, we act unwisely and immaturely. Greaves and MacAskill argue that we should overcome this immaturity by accepting the view that it is of immense moral importance to ensure that the far future goes well.

Why should we think that the far future matters so much, morally speaking? A point emphasized by Greaves and MacAskill (2025: 21–24) is how vast the future is. Many more people will live in the future than are alive today. The potential for doing future good is correspondingly large.

A similar point can be made by applying utilitarianism to questions of intergenerational ethics. A simple version of utilitarianism says that we should act in ways that maximize total welfare from an impartial perspective (without favouring our own welfare or that of those close to us over the welfare of more distant others). As economists point out, investments in the future yield positive returns. Today's generation can increase aggregate welfare across generations by saving or investing resources for the use of future generations, rather than using these resources themselves (Roser and Seidel 2016: 65). So, utilitarianism, at least in its simple version, implies that we should invest in the far future.

In the following, by the 'utilitarian view', I refer to the view that we have a moral duty to make the far future go best from an impartial perspective. If the utilitarian view is correct, this has radical implications. The utilitarian view requires us to place much more weight on far-future effects of actions than implied by common-sense morality, and to change moral practices to focus on benefitting far-future over present generations.

In this chapter, I argue that key deontological commitments are in tension with the claim that we have duties to make the far future go best. Moreover, this tension cannot be resolved simply by restricting the scope of the utilitarian view to cases in which actions do not violate deontological constraints, and are not rendered permissible by deontological prerogatives.

Charlotte Franziska Unruh, *Against a Moral Duty to Make the Future Go Best* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Charlotte Franziska Unruh 2025.
DOI: 10.1093/9780191979972.003.0008

I begin by distinguishing the utilitarian view from similar 'longtermist' views (section 2). I then describe two problems with the utilitarian view: it is implausibly demanding, and it can justify doing near-future harm for the greater (far-future) good (section 3). I examine whether proponents of the utilitarian view can solve these problems by accepting deontological constraints and prerogatives as restrictions to their view. My conclusion is sceptical. I argue that constrained utilitarianism remains in tension with deontological commitments in at least two ways. Deontological duties of beneficence are not sensitive to the stakes in the way required by the utilitarian view (section 4). Moreover, we have special duties, such as duties to those whom we have harmed in the past, which can override duties of beneficence (section 5). I suggest that an account of our special relationship to our descendants, and an account of generational sovereignty, should be key elements of a distinctly deontological approach to intergenerational ethics.

2 The utilitarian view on duties towards future generations

The utilitarian view says that we should maximize welfare across generations. It follows that we have extremely demanding duties to save for the future.[1]

One way to make the view less demanding is by discounting future outcomes, such that welfare generated in the future counts for less than welfare today. Discounting is commonly practised in economics (Roser and Seidel 2016: 67), but seems hard to justify philosophically if future people matter morally as much as present people (Greaves and MacAskill 2025: 32).

A view that is relevantly similar to utilitarianism about future generations is longtermism. According to Greaves and MacAskill (2025: 42), longtermism is the view that 'impact on the far future is the most important feature of our actions today'.[2] Greaves and MacAskill claim that the far future is vast in size and that we can predictably affect the far future, for example by investing in pandemic preparedness or in safe development of artificial intelligence. When facing choices about where to donate or how to allocate resources, investing in the far future will do much more overall good than benefitting present people. According to Greaves and MacAskill, given the stakes, we should do what is best for the far future. The conclusion of utilitarian and longtermist arguments is similar: from a moral point of view, we should do what is best for far-future generations.

The longtermist argument can be strengthened by defending the view that bringing welfare subjects into existence can benefit them. On this view, we have strong moral duties to reduce the risk of human extinction, and to expand human civilization, for example by

[1] Gosseries describes positive returns on investments as follows: 'Giving up the consumption of part of our capital today may enable us—provided it is wisely invested—to consume much more of that capital at some more or less distant future time. Consider a bag of seeds, part of which could be either consumed immediately or sown so as to multiply its volume. If you are a utilitarian, savings (in generational terms) are not just authorised; they are required since the goal is to maximise the size of the intergenerational welfare pie' (Gosseries 2008: 65–6).

[2] More precisely, Greaves and MacAskill call this view *strong longtermism*, which in its deontic version says that '[i]n the most important decision situations facing agents today, (i) One ought to choose an option that is near-best for the far future, (ii) One ought to choose an option that delivers much larger benefits in the far future than in the near future' (2025: 39). Strong longtermism stands in contrast with a weak version of longtermism, according to which we should be 'particularly concerned with' ensuring that the far future goes well (Greaves and MacAskill 2025: 17). Much depends on how 'particularly' is understood. If long-term considerations often trump near-term considerations on weak longtermism, then my arguments also apply to weak versions of longtermism.

settling space or even creating artificial sentient beings (Greaves and MacAskill 2025: 22). However, Greaves and MacAskill (2025: 32) claim that the case for longtermism does not depend on the view that bringing welfare subjects into existence can benefit these subjects: the future is big enough to give rise to demanding duties to ensure that the far future goes best, even when taking into consideration only future people that will certainly exist.

Giving up the assumption that we should add happy people to the world seems to move longtermism closer to common-sense views on caring about the future (see Setiya 2022). At least, it seems to move longtermism closer to the utilitarian view, which gives rise to a moral duty to maximize the 'intergenerational welfare pie' (Gosseries 2008: 66). In the following, I will focus on the utilitarian view.

My criticism adds to challenges for and criticisms of longtermist views that have been developed elsewhere. For example, it is controversial whether the extinction of humanity would be morally problematic (see Finneron-Burns 2017); whether moral value can be measured, compared, and predicted (see Tarsney 2023); whether we should discount future outcomes (Purves 2016; Mogensen 2022); whether we can aggregate future benefits and how strong resulting claims are (Curran 2025; Heikkinen 2022); whether benefitting the far future comes apart in practice from benefitting the near future;[3] whether we can possibly predict the effects of our actions in the far future given great uncertainty (Uzan 2022); and whether we should ignore low-probability events (but see Wilkinson 2022). Moreover, insofar as longtermism employs reasoning familiar from effective altruism, it might be vulnerable to criticisms of effective altruism such as the charge of undervaluing care, historic injustice, and inequality (Srinivasan 2015). Longtermism has also been argued to divert attention from current suffering, and justify technologies and policies that deepen power asymmetries between the powerful and the poor and marginalized (Torres 2021).

3 The case for constraining the utilitarian view

The utilitarian view on duties towards future generations fails to take seriously the separateness of persons in intergenerational contexts. In doing so, the utilitarian view generates overly demanding moral requirements, and implausibly justifies doing harm to present generations for greater future good. These criticisms are familiar from long-standing debates between non-consequentialists and consequentialists (Alexander and Moore 2021).

Nonetheless, the criticisms deserve repeating, for at least two reasons. Spelling out deontological objections to utilitarian views on duties towards future generations helps to clarify the grounds of deontological resistance against longtermist conclusions. Moreover, and importantly, understanding the nature of deontological duties to future generations helps to develop, in turn, positive deontological views on our duties towards future generations.

[3] In the words of Thorstad (forthcoming), the idea here is that one might reject '*swamping longtermism*', which says that the available far-future benefits are much greater than the available near-future benefits, and accept '*converging longtermism*', which says that actions that benefit near-future generations most also benefit far-future generations most.

Let us go back to the analogy in the introduction. Is humanity really like a lazy teenager? Upon closer inspection, the analogy seems dubious. Planning one's future is very different from planning the future of humanity. When teenagers forego immediate pleasure for hard study, they do so because they think (or are told) that studying now will benefit them later. The teenagers envision their lives and make decisions based on how they want to shape their future.

However, the situation is different when present people consider the future of humanity. Humanity is not a person, and not a subject of prudential value. There is no person who will experience all the different experiences that humanity will experience.

This point is familiar in ethics. It has famously been called the *separateness of persons* by Rawls (1971). Utilitarian principles of distributive justice distribute welfare or other goods such that the total or average amount of welfare received is maximized. The objection from a non-consequentialist perspective is that in aggregating welfare gains and losses across individuals from an impartial perspective, utilitarian principles ignore the separateness of individuals.

Maximizing overall welfare can require excessive sacrifices from some to the benefit of others. Consequently, an often-noted worry about utilitarianism is that it is very demanding. In an intergenerational context, the vast size of the future makes this worry more pressing (Gosseries 2008: 65). Consider

> *Sacrifice*: We can use our resources to prevent harm to present people, or we can use these resources to prevent harm to far-future generations. Many more people will live in the far future than are alive now. So, refraining from aiding present people and saving the resources instead will benefit more people overall.

Intuitively, if the threat to present people is significant, it is permissible, even required, to use our resources to prevent harm from present people. The demandingness of utilitarian requirements seems unacceptable (Gosseries 2008: 65).

Moreover, the utilitarian view justifies doing harm in the near future, if doing so would maximize aggregate welfare across generations.[4]

> *Catastrophe*: We can now choose a policy that will cause a catastrophe in the near future. However, the policy will also moderately improve the quality of life for a great number of generations in the far future, such that the policy would greatly increase overall welfare across generations.

It seems clearly impermissible to cause the catastrophe, even if the harm that the catastrophe causes is outweighed by the benefit caused in the far future.

Deontological moral theories can avoid these implausible implications. According to deontological theories, it is not always required to do what brings about the best consequences from an impartial point of view. Duties of beneficence are limited. For example, deontologists hold that failing to donate money to charity is often permissible. Similarly,

[4] This point is also made by Setiya, who notes that utilitarianism implies that '[w]e must sacrifice everything for the greater good' (Setiya 2022).

present and near-future generations are permitted to refrain from making extensive sacrifices in *Sacrifice*.

Moreover, according to deontological moral theories, it can be impermissible to do what brings about the best consequences. Our behaviour is subject to stringent moral constraints. For example, it is wrong to kill one person to save the lives of five others. Similarly, present and near future generations are not permitted to choose the catastrophic policy in *Catastrophe*.

In this way, deontological prerogatives protect individual agents from demanding moral duties, and deontological constraints protect individuals from harmful influences (Woollard 2015).[5]

I suggest that constraints and prerogatives can, moreover, protect the sovereignty of generations: their freedom to make their own decisions.[6] Constraints against doing harm protect generations from external restriction of their options (e.g., by prohibiting earlier generations from polluting the environment with adverse effects on future people's access to resources). Prerogatives to allow harm protect the ability of future generations to make their own choices (e.g., by allowing the use of resources for present consumption rather than saving for the future).

Now, proponents of the utilitarian view might argue that they can avoid implausible implications in *Catastrophe* and *Sacrifice*. They might try to do so by restricting their view to cases in which there are no relevant constraints and prerogatives.[7] According to this *constrained utilitarian view*, we should make the future go best, whenever doing so would not violate serious constraints or require excessive sacrifices.

In the following, I consider an argument given by Greaves and MacAskill for a constrained version of longtermism. Here is their *Stakes Sensitivity Argument*:

> (P1) When the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor, one ought to choose a near-best option.
>
> (P2) In the most important decision situations facing agents today,[8] the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor.

[5] Woollard (2015) develops and defends the view that constraints and prerogatives protect individual agents from imposition. Following Kamm (1996: 24–5), Woollard distinguishes two kinds of impositions. First, agents need protection from causal imposition. Causal imposition means performing a behaviour that is part of a causal sequence leading to a harm B suffers (Woollard 2015: 101). Second, agents need protection from normative imposition. Normative imposition means being under requirements to put our bodies at someone else's use (Woollard 2015: 101). On Woollard's view, constraints against doing harm protect agents from causal imposition, and prerogatives for allowing harm protect agents from normative imposition.

[6] The concept of generational sovereignty received renewed attention in the discussion on short-termism in politics. A central question is what it means to be free from the authority of other generations. Here, Gosseries develops a threefold distinction of increasingly broad concepts of generational sovereignty: jurisdictional sovereignty, political sovereignty, and effective sovereignty (which includes access to resources) (Gosseries 2016: 101–2). My understanding of sovereignty as the ability to make one's own decisions draws on Enoch's (2022) work on autonomy as sovereignty.

[7] Greaves and MacAskill signal their sympathy to such constraints in a footnote. Asking if their arguments might 'justify atrocities in the name of the long-term good', they answer no, because 'for such serious side-constraints, something closer to absolutism or near-absolutism becomes much more plausible' (Greaves and MacAskill 2025: 41n.38).

[8] Greaves and MacAskill discuss philanthropic giving, career choices, and policy choices.

(C) So, in the most important decision situations facing agents today, one ought to choose a near-best option. (Greaves and MacAskill 2025: 39)

The first premise of the *Stakes Sensitivity Argument* implies that plausible forms of non-consequentialism will be sensitive to the stakes. If there is a lot at stake, we have duties to do what is best, at least when no significant constraints or prerogatives are involved.

4 Duties to benefit without duties to make the future go best

The first premise of the *Stakes Sensitivity Argument* characterizes the duty of beneficence as a general duty which is, in principle, unlimited. On this view, demands of beneficence always generate moral duties to aid, unless these duties are outweighed by a prerogative.[9]

Deontologists can reject this characterization of beneficence. They can hold, against the first premise, that we have no general duty to make things go best from an impartial perspective. The deontological duty of beneficence is limited in principle, or, as Wiggins puts it, 'by its true nature' (Wiggins 2000: 191).[10]

But what, one might ask, *do* deontological duties of beneficence demand, then? As I understand the view that the duty of beneficence is limited, prerogatives do not serve to protect agents from the demands of a general and unlimited duty to do good. Rather, prerogatives allow moral agents to exercise latitude in aspects of their lives that have to do with aiding others.

To understand what it means for a duty of beneficence to be limited in this sense, consider Kant's (2019) famous view that duties of beneficence are imperfect. Imperfect duties are duties that we must fulfil often enough. We should be sensitive to the needs of others, but how, when, and where we discharge our duty to aid is up to us.

To make the concept of imperfect duties more precise, Hanser (2014) appeals to the distinction between the *justifying* strength of reasons and the *requiring* strength of reasons. Reasons that have requiring strength give rise to (*pro tanto*) duties, reasons that have only justifying strength do not. We need justification for failing to act in line with requiring reasons, but not for failing to act in line with only justifying reasons (Hanser 2014: 308). Based on this distinction, Hanser gives the following account of imperfect duties: 'an imperfect duty is a second-order reason, with requiring force, to act in accordance with first-order justifying reasons of a certain kind often enough over time' (Hanser 2014: 309).

If deontologists understand duties of beneficence as imperfect duties in this sense, then they can reject the first premise of the *Stakes Sensitivity Argument*. On the view that duties of beneficence are imperfect, agents should aid others often enough. However, there is no general duty to choose an option that is best for others, even when choosing this option is not particularly costly for agents.

[9] A view which says that we have requiring reason to save others whenever we can do so, which might be outweighed by permissions to refrain from saving, is defended by Pummer (2023).

[10] Wiggins argues against Ross's *prima facie* duty of beneficence, and develops a view on duties of beneficence as 'a schema that generates countless more specific [...] duties most of which arise (often in a supererogatory or non-mandatory manner) from the agent's historic situation, arise from who he is, arise from who his putative beneficiary is, or arise from what goals he has already committed himself to promoting' (Wiggins 2000: 190). I thank John Tasioulas for drawing my attention to Wiggins's paper.

One might object that the view that duties to aid are imperfect leaves open the possibility that *if* we aid, *then* we should aid where the stakes are highest. Imagine a scenario in which agents are deciding whether to use a certain amount of money to benefit either near-future or far-future people. Given the stakes, one might argue that the justifying reason to benefit far-future people is stronger than the justifying reason to benefit near-future people. It seems that the second-order duty to aid should tell people to spend their money in a way that is favoured by the strongest justifying reasons to aid, which in this case is to benefit the far future.[11]

However, the second-order duty to aid need not be understood as a duty to act in accordance with the strongest justifying reasons. Consider, e.g., the view that the second-order duty to aid is a duty to 'do one's share' (Hanser 2014: 313). Ensuring that we do our share in aiding others does not require ensuring that we maximize aid.[12][13]

Moreover, the strength of justifying reasons to aid plausibly depends on factors such as our commitments, intentions, and relationships to present and future people, in addition to or instead of the number of interests at stake.

Another objection on behalf of proponents of the *Stakes Sensitivity Argument* might be that imperfect duties can be construed such that they are compatible with the first premise of the argument. Hanser himself points out that there is a second way in which imperfect duties can be understood. On this second view, first-order reasons to aid are requiring reasons, but second-order permissions can make non-compliance permissible. In Hanser's words: 'an agent has an imperfect duty to act in a certain way whenever he has both (i) a first-order reason, with requiring force, to act in that way, and (ii) a second-order permission to ignore first-order requiring reasons of that kind from time to time' (Hanser 2014: 311).

A key difference between the two views is who bears the justificatory burden. On the first view, the person who imposes a requirement needs to show that agents have a requiring reason to act in accordance with justifying reasons. On the second view, agents need to show that they have permission for failing to act in accordance with requiring reasons.

Deontologists should reject the second view. This is because bearing the justificatory burden is, in itself, an imposition. Prerogatives should protect agents from demanding moral requirements. On the first view, prerogatives provide this protection, as agents do not bear the justificatory burden. On the second view, prerogatives do not provide this protection.[14]

[11] I thank an anonymous reviewer for raising this objection.

[12] Importantly, what gets shared here is the burden of aiding others enough, not the burden to maximize welfare. More demanding duties of beneficence cannot appeal to the 'fair shares' view (see Mogensen 2021).

[13] Sinclair (2018) makes similar points in response to Pummer's (2016) and Horton's (2017) defences of conditional duties to donate effectively. On the conditional duties view, we are not required to aid more, but we are required to aid others efficiently once we have decided to aid. Sinclair argues that Pummer and Horton, despite conceding non-consequentialist options, still use a consequentialist background structure of the moral landscape, e.g., by assuming that there is an impartial duty to aid that gives rise to claims on behalf of potential beneficiaries (Sinclair 2018).

[14] My claim here is only that there is no *general* requirement that agents justify their behaviour whenever they can aid. This is compatible with the view that agents can sometimes be subject to such requirements. This point is similar to, and inspired by, an argument by Woollard (2018) that mothers do not have defeasible duties to benefit their children. On Woollard's (2018) account, 'defeasible' duties require agents to justify their behaviour in case of non-compliance.

A final defence of the *Stakes Sensitivity Argument* is that its conclusion should be understood in a weak sense. Even if we are not required to do what is best, it is still *better* to do more long-term good than less near-term good.[15] However, if self-sacrifice for the present generation and self-sacrifice for future generations both lie beyond the call of duty, it is unclear exactly what difference future stakes make. Saving resources for the future seems to become one of many ways in which we can discharge our imperfect duty of beneficence. This result is far less revisionary than the utilitarian view.

5 Special duties to future generations

Let us consider the view that benefitting the long-term future is one way to fulfil one's duty of beneficence. On this view, it is not required to devote one's life to making the far future go best, but it is permissible to do so. To see whether this view is correct, we should examine whether it could ever be *impermissible* to do what is best for the far future.

In the *Catastrophe* case, we have seen one way in which improving the far future could be impermissible: if doing so entails the violation of a moral constraint against doing harm.

I suggest that there is another way in which improving the far future could be impermissible: if doing so entails *failing to undo* threats of harm that we have initiated.

Importantly, I do not claim that these are the *only* relevant considerations in determining whether saving for the far future is permissible. It is plausible that duties of justice to our contemporaries can prohibit saving for the future. For example, we might have strong duties to prioritize aiding the worst off in our own generation (see Gosseries 2008: 68). Moreover, saving for the far future might sometimes objectionably interfere with the sovereignty of present or near-future generations.[16]

Now, consider that our actions can pose threats to future generations. Constraints against doing harm to future generations might tell us to refrain from overly polluting the environment, or to refrain from implementing dangerous technologies.

I argue that when we have already imposed threats of harm, and can now prevent this threat from materializing, we should prevent ourselves from doing harm.

Letting oneself do harm (i.e., failing to prevent one's previous actions from leading to harm) is harder to justify than merely failing to benefit. To illustrate, consider a case given by Hanna (2015, 678). In Hanna's case, Agent has left poisoned tea in his home, which he knows Victim will drink in ten minutes. Agent decides to drive back home and pour the tea away. On his way, Agent drives past five people who need urgent help. If Agent stops, the five will live, but Victim will drink the tea and die. If Agent drives past the five, they will die, but Victim will be saved in time. Intuitively, Agent is *required* to drive back, as Hanna suggests, 'precisely because Agent otherwise will have killed' Victim (Hanna 2015: 679). It seems even clearer that Agent is *permitted* to drive back, even if this will save the lesser number.

[15] I thank Andreas Mogensen for discussion on this point.

[16] The tension between generational autonomy and longtermism is noted by Hoffmann (2022). Ord (2020: 190–93) suggests a period of 'Long Reflection' spanning several generations to develop a strategy for developing humanity's potential. Hoffmann argues that imposing such a reflection period on intermediary generations conflicts with what Hoffmann calls 'intertemporal sovereignty'. Moreover, Ord assumes that the next generations will or should care about humanity's 'potential', which they might not do (Hoffmann 2022: 80).

How can we explain the special moral status of letting oneself do harm? Hanna considers a plausible response: We have special obligations to our past victims. This explains why letting oneself do harm is harder to justify than merely allowing harm or benefitting (Hanna 2015: 686).[17]

As I argued elsewhere (Unruh 2021), the discussion on letting oneself do harm has real-world applications: it maps onto cases in which we have imposed risks of harm, and now face a choice of either using our resources to undo (prevent, mitigate, or compensate) these harms, or using our resources to benefit the far future.

If our reasons for undoing harms can be stronger than reasons to prevent harms, then it can be required to focus on undoing our harms, rather than make the future go best. This gives rise to an objection to the first premise of the *Stakes Sensitivity Argument*: beyond constraints against doing harm, duties to undo harm can limit duties to benefit.

One might object to this argument by appealing to the *non-identity problem* (Parfit 1984). The idea is that we cannot undo harms to future generations without changing the identity of those who will be affected. That is, undoing our harms will not benefit those who would have been affected by our previous actions (rather, it prevents these people from coming into existence in the first place). In this way, non-identity might weaken the reason for undoing harm to future generations.

In response, the non-identity problem also applies to duties to benefit, since actions that benefit future people might change the identity of these future people (Unruh 2020). More generally, if non-identity matters morally, then it would seem to follow that we generally have stronger duties towards present generations than towards far-future generations, whose identity depends on what we do today. Such a result would weaken, not strengthen, the utilitarian view.[18]

6 Conclusion

I have argued that deontologists should reject the utilitarian view, even in its constrained version. Humanity today is not like a teenager planning for his later life. Limits to duties of beneficence ensure that we are not subject to moral requirements to make the future of humanity go best. Also, it can be impermissible to do what makes the future go best when this harms or limits the sovereignty of present or future generations.

What emerges is that duties towards future generations, from a deontological perspective, are multi-layered and complex. The relations we stand in to present, near- and far-future people, the ways in which our behaviour can affect them, and other moral commitments can make a moral difference to the content of our duties towards future people. A deontological approach to intergenerational ethics should explore the nature and moral

[17] Hanna raises concerns for this response: it seemingly cannot account for cases in which we can prevent ourselves from doing future harm (Hanna 2015: 687). For a response to this concern, see (Unruh 2021: 383).

[18] For an argument that the non-identity problem can give rise to an objection to the *High Stakes Argument,* see Mogensen (2019). Purves (2016) argues that non-identity can justify discounting future welfare. A different version of the objection I discuss says that 'undoing' our own harms is likely to benefit people different from those we have harmed. For example, Stefansson (2022) argues that offsetting carbon emissions is unlikely to prevent the very same harm that our emissions would have caused.

significance of the relationship between, and the moral implications of, the sovereignty of present and future generations.

My argument, if correct, has practical implications. Emerging technologies will likely have significant effects on the far future. These technologies also shape the lives of people alive today and in the near future. There is an urgent need to address adverse effects of new technologies that are materializing today, such as large-scale surveillance and inequality. We should certainly not neglect considering the potential far-future effects of new technologies. But, if I am correct, then the mere size of the far future, by itself, does not give us reason to prioritize far-future over more immediate concerns. From a deontological perspective, it might well be that we have strong moral duties to future generations. But a general duty to make the future go best is not among them.

Acknowledgements

I thank an anonymous referee, Jonathan Hoffmann, and Elad Uzan for excellent comments on previous drafts of this chapter. I also thank audiences at the MCMP at the Ludwig-Maximilians-University Munich, the Global Priorities Institute, the Legal Priorities Lab, and Nova University of Lisbon for helpful comments and discussion. I am grateful to André Santos Campos and members of the research project "Present Democracy for Future Generations" at Nova University of Lisbon for stimulating discussion. This work benefitted greatly from my participation in the Early Career Conference Programme 2021 at the Global Priorities Institute, University of Oxford. I thank the Global Priorities Institute for hosting me, and especially Andreas Mogensen for helpful discussion.

References

Alexander, L. and Moore, M. (2021), 'Deontological Ethics', in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2021 (Metaphysics Research Lab, Stanford University), https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/. Accessed 27 December 2022.

Curran, E. (2025), 'Longtermism and Aggregation', in Philosophy and Phenomenological Research 110/3: 1137–1151.

Enoch, D. (2022), 'Autonomy as Non-Alienation, Autonomy as Sovereignty, and Politics', in *Journal of Political Philosophy* 30/2: 143–65.

Finneron-Burns, E. (2017), 'What's Wrong with Human Extinction?', in *Canadian Journal of Philosophy* 47/2–3: 327–43.

Gosseries, A. (2008), 'Theories of Intergenerational Justice: A Synopsis', in *S.A.P.I.EN.S. Surveys and Perspectives Integrating Environment and Society* 1/1: 61–71.

Gosseries, A. (2016), 'Generational Sovereignty', in I. González-Ricoy and A. Gosseries (eds.), *Institutions For Future Generations* (Oxford University Press), 98–114.

Greaves, H. and MacAskill, W. (2025). 'The Case for Strong Longtermism', in H. Greaves, J. Barrett and D. Thorstad (eds.), Essays on Longtermism (Oxford University Press). 17–43.

Hanna, J. (2015), 'Doing, Allowing, and the Moral Relevance of the Past', in *Journal of Moral Philosophy* 12/6: 677–98.

Hanser, M. (2014), 'Imperfect Aiding', in S. Luper (ed.), *The Cambridge Companion to Life and Death* (Cambridge University Press), 300–15.

Heikkinen, K. (2022), 'Strong Longtermism and the Challenge from Anti-Aggregative Moral Views', GPI Working Paper No. 5-2022 (Global Priorities Institute, Oxford University), 1–29, https://globalprioritiesinstitute.org/karri-heikkinen-strong-longtermism-and-the-challenge-from-anti-aggregative-moral-views/. Accessed 13 August 2022.

Hoffmann, J. (2022), *Democracy, Justice, & the Long Term – Designing Institutions for the Future*, PhD thesis, University of Warwick.
Horton, J. (2017), 'The All or Nothing Problem', in *Journal of Philosophy* 114/2: 94–104.
Kamm, F. (1996), *Morality, Mortality: Rights, Duties, and Status*, vol. 2 (Oxford University Press).
Kant, I. (2019), *Grundlegung zur Metaphysik der Sitten,* T. Valentiner (ed.), Reclams Universal-Bibliothek, Nr. 4507 (Reclam).
Mogensen, A. L. (2019), 'Staking Our Future: Deontic Longtermism and the Non-Identity Problem', GPI Working Paper No. 9-2019 (Global Priorities Institute, Oxford University), 1–32, https://globalprioritiesinstitute.org/wp-content/uploads/2020/Andreas_Mogensen_staking_our_future.pdf. Accessed 11 June 2021.
Mogensen, A. L. (2021), 'Moral Demands and the Far Future', in *Philosophy and Phenomenological Research* 103/3: 567–85.
Mogensen, A. L. (2022), 'The Only Ethical Argument for Positive δ? Partiality and Pure Time Preference', in *Philosophical Studies* 179: 2731–50.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books).
Parfit, D. (1984), *Reasons and Persons* (Oxford University Press).
Pummer, T. (2016), 'Whether and Where to Give', in *Philosophy and Public Affairs* 44/1: 77–95.
Pummer, T. (2023), *The Rules of Rescue: Cost, Distance, and Effective Altruism* (Oxford University Press).
Purves, D. (2016), 'The Case for Discounting the Future', in *Ethics, Policy & Environment* 19/2: 213–30.
Rawls, J. (1971), *A Theory of Justice* (Harvard University Press).
Roser, D. and Seidel, C. (2016), *Climate Justice: An Introduction* (Routledge).
Setiya, K. (2022), 'The New Moral Mathematics', in *Boston Review*, 15 August 2022, https://www.bostonreview.net/articles/the-new-moral-mathematics/. Accessed 28 November 2022.
Sinclair, T. (2018), 'Are We Conditionally Obligated to Be Effective Altruists?', in *Philosophy & Public Affairs* 46/1: 36–59.
Srinivasan, A. (2015), 'Stop the Robot Apocalypse', in *London Review of Books* 37/18, 24 September 2015.
Stefansson, H. O. (2022), 'Should I Offset or Should I Do More Good?', in *Ethics, Policy and Environment* 25/3: 225–41.
Tarsney, C. (2023), 'The Epistemic Challenge to Longtermism', in *Synthese* 201/6: 195.
Thorstad, D. (forthcoming), 'The Scope of Longtermism', in *Australasian Journal of Philosophy*.
Torres, P. (2021), 'The Dangerous Ideas of "Longtermism" and "Existential Risk"'. *Current Affairs*, 28 July 2021, https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk. Accessed 3 June 2022.
Unruh, C. F. (2020), 'Can We Benefit in Non-Identity Cases?', in *Intergenerational Justice Review* 5/2: 49–50.
Unruh, C. F. (2021), 'Letting Climate Change', in *Journal of the American Philosophical Association* 7/3: 368–86.
Uzan, E. (2022), 'Moral Mathematics: Subjecting the Problem of Ethics to the Cool Quantifications of Logic and Probability Can Help Us to Be Better People', *Aeon* (Blog), 28 November 2022, https://aeon.co/essays/how-to-solve-moral-problems-with-formal-logic-and-probability. Accessed 4 December 2022.
Wiggins, D. (2000), ' "The Right and the Good" and W. D. Ross's Criticism of Consequentialism', in *Royal Institute of Philosophy Supplement* 47: 175–95.
Wilkinson, H. (2022), 'In Defense of Fanaticism', in *Ethics* 132/2: 445–77.
Woollard, F. (2015), *Doing and Allowing Harm* (Oxford University Press).
Woollard, F. (2018), 'Motherhood and Mistakes about Defeasible Duties to Benefit', in *Philosophy and Phenomenological Research* 97/1: 126–49.

9

Authenticity, Meaning, and Alienation

Reasons to Care Less about Far-Future People

Stefan Riedener

1 Introduction

What ought we to do today? Take the United States, and consider just two of their options. At the time of writing, Russia is attacking Ukraine. The Ukrainians must defend themselves, accommodate refugees, and start to rebuild their country. The US could help them by sending weapons or money or welcoming escapees. At the same time, artificial intelligence (AI) technology is progressing. There's a risk that an AI will go out of control, and cause enormous harm. The US could slightly reduce that risk by investing in AI safety. What ought they to do? According to longtermism (as I'll use the concept), in situations like these, what we ought to do depends almost exclusively on the very long-term effects of our actions. What matters is how the US's decision affects the next millions of years. It matters little, in itself, how it affects the coming year, century, or even millennium.

The standard argument for this view builds on four simple premises.[1] First, the long-term future concerns the fate of an enormous number of people.[2] Humanity might colonize the universe, survive for billions of years, and live lives of unimaginable quality. Or we might soon go extinct, or remain a small group of miserable souls leading a wretched form of existence. The difference might concern, say, a septillion (10^{24}) people. Second, in general, far-future people should matter to us just as much as contemporaries do. Consider spatial distance. If we can help someone, it shouldn't matter to us whether they're here or 10,000 kilometers away. The same is true for temporal distance. If we can do someone good, it shouldn't matter to us whether they live now or in a million years. Third, we can affect the existence and quality of life of these far-future people. Of course, we can't predict long-term effects with any confidence. But in very many decision situations, at least, certain actions seem to have a *somewhat* higher probability than others of positively impacting the long-term future. And even just a one-in-a-billion ($1/10^9$) chance to affect a septillion people, say, in expectation still affects a quadrillion (10^{15}). So even minuscule increments in probabilities *ex ante* affect *a lot*. Fourth, typically, if we can expectably affect the existence and life-quality of billions or quadrillions of people, we simply ought to do it. Our reasons to do

[1] Roughly, this argument is given e.g. in MacAskill (2022: ch. 1), although MacAskill omits the fourth premise, and thus doesn't draw a similarly strong conclusion. For a similar argument, with a similar conclusion, see e.g. Greaves and MacAskill (2021) or Greaves, MacAskill, and Thornley (2021).

[2] I'll follow much of the existing literature by making two simplifications. First, I'll set aside non-human animals, although they clearly matter too. Second, I'll call our far-future descendants 'people', although they needn't be humans (or *Homo sapiens*) by any means.

Stefan Riedener, *Authenticity, Meaning, and Alienation* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Stefan Riedener 2025. DOI: 10.1093/9780191979972.003.0009

so trump just about any other consideration. They're more important than our reasons to affect the existence and life-quality of fewer people—or to keep our promises, show loyalty to our friends, or even to refrain from harming.

Prima facie, the upshot of this simple reasoning is (what I'll call) longtermism: typically, we simply ought to do what's expectably best in the long term. All other considerations—in particular, our effects on the coming century or so—are largely irrelevant in themselves, or relevant only as tie-breakers when all of our options are *ex ante* equally good in the longer term.[3] Take the US decision again. According to many experts, AI poses one of our greatest extinction risks. This risk can be reduced. Yet attempts at doing so are still very sparse. So investments in AI safety seem among the best longtermist interventions.[4] By contrast, housing Ukrainian refugees or building up Ukrainian churches doesn't seem to affect the long-term prospect much. And while military aid might have long-term benefits in deterring aggressive warfare, it also involves a real risk of inducing a catastrophic escalation.[5] So the long-term benefits of helping Ukraine seem decidedly less clear. Presumably, according to longtermism, the US ought to invest in AI safety. And in any case, whatever the ultimate upshot, the 43 million Ukrainians themselves are practically irrelevant for the decision. The US's dominating concern should be to marginally increase the chance that in a million years, in remote corners of space, quintillions of our descendants enjoy fantastic lives.

This is an extremely revisionist idea, to put it mildly. If the above assumptions are right, it would often imply we should take radically different actions than we're currently taking. And it would just about always imply we should base our actions on radically different considerations than we're currently doing.[6] Notably, the just-mentioned argument isn't simply utilitarian. It applies on any view that gives relevantly non-trivial weight to promoting expected total welfare. Still, it involves a number of controversial assumptions. Some people might oppose population ethical presuppositions of the argument as I presented it—e.g. the idea that we have strong reasons to bring people into existence. Others might object to decision-theoretic assumptions—e.g. the idea that you ever ought to prefer such tiny-probability high-payoff options. Still others might oppose empirical premises—e.g. the idea that we aren't entirely clueless about our long-term effects. In this chapter, I'll grant the longtermist all of these assumptions. For what it's worth, I tend to think they're either plausible, not ultimately essential to the argument, or not pointing to the fundamental problem about it. So instead, I focus on the second premise of the above argument—the thought that we should care just as much about far-future people as about our contemporaries. This is a

[3] There's currently no consensus about what exactly 'longtermism' means. The argument I've just sketched is very natural. I think it's not uncommon. And it's the conclusion of this argument that interests me here. So, I use 'longtermism' to denote this conclusion. Some people use it to denote the weaker view that the long-term future should be *a* key moral priority of our time; they use 'strong (deontic) longtermism' to denote the view I have in mind (see e.g. Greaves and MacAskill 2021; MacAskill 2022: 4). Depending on what it means to be 'a key priority', I think my arguments are compatible with this weaker view. So in this alternative terminology, I don't object to longtermism, but to strong (deontic) longtermism only.

[4] See e.g. Ord (2020: chs. 5 and 6).

[5] This has been a guiding concern of the White House in deciding what help to provide. See e.g. Sanger et al. (2022).

[6] To illustrate this with our initial example: as of December 2023, US aid to Ukraine totaled $68 billion, and the White House has asked Congress for another $37.7 billion; in contrast, recent estimates suggest that *global* annual spending on reducing existential risks from AI is between $10 and $50 *million*. See e.g. Ord (2020: 58 fn. 55) and Cancian (2022).

very revisionist idea itself. As a matter of fact, we care much more about people in Ukraine, say, than about our distant descendants. I want to offer a philosophical justification of our unequal concern, and explore how this affects the longtermist argument.

To that end, I'll first consider existing understandings of 'temporal discounting', and argue that on these interpretations, our discounting seems either unwarranted or insufficient to block the argument (section 2). I'll then offer two alternative reasons to care less about far-future people. The first consideration concerns our *reactive* relation to far-future goods: the way in which we respond to, or appreciate, those goods. I'll argue that our caring equally much about far-future people would be problematically inauthentic (section 3). The second consideration concerns our *proactive* relation to far-future goods: the way in which we potentially bring those goods about. I'll argue that the way in which we could do good to far-future people contributes much less meaning to our lives than the way in which we can benefit contemporaries (section 4). I'll end by discussing objections (section 5). If I'm right, my arguments are compatible with the view that we should be doing much more for the far future than we're currently doing. Still, they imply that in many situations, we ought to act for (roughly[7]) contemporaries' sakes, even if longtermist actions would expectably do much more good overall.

2 Preliminaries and premises

The idea that we should care less about future people has been discussed a lot. The simplest interpretation of it is in terms of temporal discounting, in the form of a 'pure time preference'. On this interpretation, temporal distance itself is fundamentally normatively relevant: if you can provide two goods x and y, but y would occur in the more distant future, this itself is a reason to prefer x. In the long term, such discounting—even by small factors—can have dramatic effects. If goods at any time are 99.99% as important as goods a year earlier, say, our interests are more than 10^{43} times as important as those of people in a million years. This would arguably undercut the far future's overwhelming importance.

However, few people find such a pure time preference compelling. Temporal distance itself just doesn't seem fundamentally normatively relevant. Here's one way to see this. Suppose that, without anyone realizing anything, the universe will freeze for a millennium at midnight today. Subjectively, the night will feel as short as any. But actually, the macroscopic world will stand still for a thousand years. It doesn't seem that you'd then have less reason today to try to benefit people 'tomorrow'—e.g. to buy your partner a present for their upcoming birthday. Temporal distance, *in itself*, just doesn't seem to fundamentally matter.[8]

If anything, it's other properties which matter intrinsically, perhaps generally *correlate* with temporal proximity, and allow us to care less about future people. The most salient idea is personal relationships. Intuitively, you ought to care more about someone the closer your relationship is to them. So perhaps we generally ought to care less about far-future

[7] In what follows, by 'contemporaries', I mean people living now *or* in the nearer future (e.g. the next couple of hundred years or so). By 'far-future people', I mean people living in the very distant future (e.g. 10,000 or 100,000 years from now).

[8] For prominent critical discussions of pure time preference, see e.g. Sidgwick (1907: 381, 414), Ramsey (1928), Parfit (1984: Appendix F), Broome (1991, sec. 8.3), or Rawls (1999: sec. 45). For a helpful survey, see Greaves (2017).

people because we have a weaker relation to them than we have towards our contemporaries. Mogensen (2022) provides one way to flesh this out. Intuitively, you ought to care more about someone the more closely you're *related* to them. Now consider the perspective of humanity's current generation as a whole. Perhaps we're less closely related to future generations the further away in time they are—just like you're less closely related to your great-grandchildren than to your grandchildren, say. And so perhaps our current generation as a whole ought to care more about near future people than about people further away. Mogensen himself doesn't apply this idea to longtermism, or claim that it undermines the above rationale. But one might think it does.

However, I'm skeptical. One might doubt that kinship is normatively important in the first place, that it makes sense to speak of entire generations as being related to each other, or that this collective perspective is sufficiently relevant to the overall longtermist agenda.[9] But even setting these worries aside: the idea simply doesn't seem to allow us to discount enough. Note that you'll never be less closely related to any of your descendants than you are to a completely unrelated contemporary. So (as Mogensen emphasizes) in light of your kinship, you never ought to care less about future people than about such unrelated people of the present. But intuitively, we ought to care very considerably about the latter. We ought to care greatly about people in Ukraine, say, even if we aren't related to them.[10] The prospect of affecting a quadrillion such unrelated contemporaries would still seem more pressing than almost anything else. If we ought to care no less than this about far-future people, they still seem dominantly important.

A similar issue arises with other salient accounts. Consider gratitude, compensation, or friendship. Intuitively, you ought to care more about people who've benefitted you, people you've harmed, or people who are your friends, than about strangers with whom you've never interacted in any way. And plausibly, in general (though not with the same kind of necessity as in the case of kinship), we have stronger relationships of these kinds to contemporaries than to far-future people. In light of the global economy, very many contemporary people have benefitted us in some way.[11] We have harmed very many people.[12] And societies like the US and Ukraine might be regarded as friends of some kind. So plausibly, in light of these considerations, we generally ought to care more about contemporaries than about people in the far future. But again, you ought to care considerably about strangers, to whom you owe no obligations of gratitude, compensation, or friendship. If you can affect a quadrillion such strangers, that seems to trump almost anything. These kinds of 'discounting' for special relationships don't seem to block our argument either.[13]

[9] Why focus on entire generations in the first place? The reason is that it's not clear whether, as an individual, you're less related to a random person in 10,000 years, say, than to a random person of the present.

[10] Of course, given our common origins, you're somewhat related to every human being. To eliminate this point, suppose half of humanity descended from (perfectly human-like) aliens who arrived here long ago, and you aren't related to some contemporaries at all. You still ought to care considerably about them.

[11] For an illustration of this, see e.g. Jacobs (2018). I thank Adam Lovett for this reference.

[12] For a famous defense of this idea, see e.g. Pogge (2002).

[13] Something similar applies to other possible justifications of discounting. Parfit discusses whether you may care less about your future selves insofar as you're less psychologically connected to them (1984: 313–317). We might turn this into an argument for discounting future generations, by considering the prudential perspective of humanity as a whole. However, your future self will plausibly never be less connected to you than a complete stranger. And you ought to care considerably about strangers. For other justifications of temporal discounting in terms of relationships, see e.g. De-Shalit (1994: 85) or Lloyd (2021).

So we face a challenge. The simplest interpretation of the relevance of temporal distance is implausible, and more complex interpretations don't seem to get us near enough discounting to undermine the longtermist conclusion. The general problem is that considerations like kinship, or gratitude, compensation, and friendship, give us reason to care *more* about some people than about completely unrelated contemporaries. But we should care considerably about the latter. So what we need, it seems, is features of far-future people which mean we ought to care *less* about them than about such contemporary strangers. I think the two considerations I'll introduce in what follows have precisely this form. Neither of them correlates linearly, or with any metaphysical necessity, with temporal distance. But at least for contingent reasons, they're generally salient for far-future people. So in the next two sections, I'll suggest that these considerations undermine the second premise of the above argument: they imply we generally ought to care *less* about far-future people than about contemporary strangers. In section 5, I'll then suggest they also undermine the longtermist conclusion: they imply we shouldn't generally just do what's expectably best in the long term.

To make my arguments, I'll rely on two assumptions: a normative and an empirical premise. I personally think they're true. But they're controversial, and thus constitute limitations of my argument. My first, normative assumption is an

> *Objective list theory of welfare:* What makes someone's life good for them is (at least in part) its containing instances of a pluralist list of objective goods: friendships, achievements, virtues, and so on.

With this assumption I'm ruling out, in particular, simple hedonist and desire satisfactionist theories. I'm assuming that what makes your life good for you isn't always just that it's pleasant for you or satisfies some desire of yours.[14] Instead, the grounds of your welfare aren't reducible to any single welfare-good, and are at least partly independent of your attitudes towards them.[15]

My second, empirical assumption, concerns the nature of far-future people's lives. It's the

> *Assumption of unknown differences:* If far-future people exist, their lives—e.g. the kinds of relationships, projects, art, understanding, or character traits they'll have—will be radically different from ours. We're quite clueless of how exactly their lives will look, and our influence on it is enormously indirect at best.

The key thought here is simple. If cultural and technological development continues even roughly on the current trajectory, life in a million years won't be remotely like ours. This is something that advocates of longtermism have typically emphasized. For instance, Ord (2020: 22) says: 'beauty, understanding, culture, consciousness, freedom, adventure, discovery, art—our descendants would be able to take these so much further, perhaps even

[14] For helpful overviews of hedonism and desire satisfactionism, see e.g. Gregory (2015) and Heathwood (2015) respectively.

[15] For some relevant discussions of objective list theories, see e.g. Finnis (1980: chs. 3–4), Parfit (1984: Appendix I), or Fletcher (2013). Note that I'm only presupposing that objective goods are *part* of what grounds your welfare. This doesn't rule out that your subjective attitudes to these goods matter too. For an overview of such 'hybrid views', see e.g. Woodard (2015).

discovering entirely new categories of value, completely unknown to us'. Similarly, MacAskill claims that 'we can't even imagine the heights that human accomplishment might reach.'[16] And Bostrom (2008: 2) describes far-future life-quality in a similar tone: 'Beyond dreams. Beyond imagination.' This seems plausible. A million years is a very long time. Perhaps our coffee conversations won't concern the weather, the neighbors, or children, but hyper-real diophantine approximations and ultraviolet matter in nano-quark holes. Perhaps we'll be receptive to magnetic fields, and enjoy 'schmusic'-performances, in which magnets of different sizes are moved around us in different patterns. Maybe we'll have found ways of 'merging' with others, as the ultimate form of friendship or sex. Perhaps concepts like 'benevolence', 'courage', or 'compassion' will have become vain, as we're automatically reasonable and happy, and the core virtue will be 'post-humaneness'. Perhaps we'll have evolved into cyborgs or AI entirely, and spend our lives relishing the delights of the 'schmetaverse'. We have no clue, and at best an extremely indirect influence on it. Or so I assume. Let's now see how these assumptions affect our concern for far-future people.

3 Authenticity

Suppose Leah is a Zen teacher in Melbourne. She guides meditations, lectures about Zen, and organizes retreats. She got into it through her former partner. But one day she admits to herself that it's all been a fancy façade. She's never really experienced that clarity of mind in meditation, which she so often extols. She generally finds that sitting and breathing tedious. She doesn't really believe, or even think she truly understands, the mysterious Zen philosophy. If anything, it's during her modern dance classes, with her dream of becoming a professional dancer, or with an aesthetical ethics à la Nietzsche, that she feels truly at home. Perhaps it was all a sophisticated way of impressing her former partner, or indeed herself.

Intuitively, there's something problematic about Leah's life. She doesn't *really* seem to value the things around which she orients her outward actions. Her life seems inauthentic. To substantiate this thought, here's a way to define (at least a core aspect of)

> *Authenticity:* It's inauthentic for one to respond to something's value in one's action (at least in part) to the extent that one doesn't genuinely appreciate that value.

Let me explain. What is it for you to 'genuinely appreciate' something's value? To begin with, it's for you to actually care about that value, with core parts of your person, beyond your intentional action. It's for you to not just act in accordance with it, but to also have appropriate affective, conative, or cognitive responses to the value. Thus you don't genuinely appreciate the value of neighborship if you never have the faintest desire to live in proximity to others. If you can witness sexual violence without a trace of aversion, or even find pleasure in it, you don't really appreciate its wrongness. And perhaps you don't genuinely appreciate the value of non-human animals if you're certain they have no moral status. In this sense, we might say, authenticity is about staying true to *you*—about aligning your actions with the things that actually matter to you deep down.

[16] See MacAskill (2018; at 7:50).

But to genuinely appreciate something's value, it's not enough for you to *have* these responses to it. You must also have them for the right reasons. To the extent that something is valuable in virtue of certain valuable-making properties, you must care for the value in virtue of caring about these properties. This means you must have access to these grounding properties—in virtue of your intellectual understanding, physical capacities, life-experience, or whatever—and must have the relevant affective, conative, or cognitive responses to these grounding properties too. Thus, if you've lost your taste functions to COVID-19, you can't genuinely appreciate the deliciousness of that pumpkin soup. You can't really appreciate the brilliance of Parfit's *Reasons and Persons* if you have no understanding of its theses and arguments, or if you think they're all profoundly misguided, or if they all leave you entirely cold. And perhaps, if you've never loved anyone, you can't genuinely appreciate the full value of love. In this sense, we might say, authenticity is also about staying true to a value—about your actions flowing from a genuine understanding of it, rather than a superficial or accidental response to it.[17]

According to the above account, insofar as you don't genuinely appreciate something's value, it's inauthentic for you to respond to it in your action. So, if you respond to these values in such cases—if you compliment the chef on that pumpkin soup (barred of your taste functions), do activism against sexual violence (against a background of indifference about it), or publicly praise *Reasons and Persons* (while it's actually all Greek to you)—it will be inauthentic.[18] Now the account doesn't say that it's necessarily (even *pro tanto*) *morally* problematic to act inauthentically. Perhaps even if you don't genuinely appreciate the wrongness of sexual violence, say, there needn't be anything morally bad about protesting against it. Still, intuitively, the fact that an action would be inauthentic for you provides a reason against it, or mitigates your reasons for it. Other things equal, if one of your actions would be more authentic than another, you have more reason to perform the former.

This seems to capture the inauthenticity of Leah's life. Leah doesn't have full access to the grounds of the value of Zen: she doesn't thoroughly know that clarity of mind, doesn't really understand its philosophy. She lacks many of the relevant affective or conative responses to that value: she doesn't unreservedly enjoy the Zen practice, and feels no real admiration for it all. To that extent, the way in which she nonetheless adheres to that value in practice—her teaching, lecturing, or extolling—is inauthentic. *Ceteris paribus*, Leah thus seems to have more reason to do modern dance rather than Zen.

If all of this is right, this ideal of authenticity will affect the longtermist premises. In light of the objective list theory, in caring about far-future welfare, you should care about far-future objective goods. But in light of our assumption of unknown differences, you can't genuinely appreciate these goods' value. Consider schmusic, say: you're not receptive to magnetic fields, and in witnessing a 'piece', you'd presumably respond with bewildered amusement instead of appropriate awe. Similarly, consider the value of partner-merging or being 'post-humane': you have no access to why these things are good, and would presumably be more appalled or left indifferent than appropriately impassioned by them. You can't really value those values. So if you take longtermist actions with the goal of promoting

[17] Some people might agree that a practical concern for a higher-order property, without appreciation of its grounds, is problematic—but doubt that it's a problem of 'authenticity'. Not much hinges on this label for my purposes. What matters is that there's something problematic about such concern.

[18] For related ideas on authenticity, or 'integrity', and for defenses of its importance, see e.g. Taylor (1992), Varga (2011), Paul (2014), Archer (2017), or Bauer (2017).

schmusic, merging, and post-humaneness, there's something inauthentic about it. It's a bit like praising a soup you haven't tasted, or like Leah's promoting Zen.

Actually, in fact, the inauthenticity of longtermist actions must cut considerably deeper. After all, it's not that we *know* the constituents of far-future welfare—like schmusic, partner-merging, or the schmetaverse—but can't appreciate these things properly. We're largely clueless about what will make far-future lives good. So it would be unreasonable to take longtermist actions with the intention to promote specific goods in particular: the probability of promoting any actually specifiable thing is overly tiny. More reasonably, we may take such actions with the intention to promote far-future welfare in the abstract, whatever it will consist in: we may try to contribute to future people *existing*, perhaps under favorable conditions, and hope that over time our descendants will have improved their form of life to whatever specific effect. In other words, our motivation for longtermist action can't be a *de re* concern with the grounds of far-future welfare. It must be a concern with such welfare as such, or *de dicto*. Compare this with contemporaries. Of course, even in donating to Oxfam, say, you can't know how exactly you'll help. But by and large, you know what makes present people's lives good: our familiar relationships, achievements, virtues, and so on. So you're dealing with known unknowns: you know that if you donate to Oxfam, someone will get something like *that*. And in virtue of caring *de re* about every item from our objective list, you can care about the uncertain prospect of benefiting someone in some such way. But that's not so with far-future people. Here you're dealing with unknown unknowns: things you currently can't even imagine. So, you can't care about the uncertain prospect of benefiting far-future people *in virtue* of caring about some list of objective goods. If you care about this prospect, it will be because you care about the sheer higher-order fact that far-future people live well. You'll try to do them good, without any idea about what this goodness might consist in.

Now, some people might think that to care about welfare *de dicto* is 'fetishistic' in a strong sense: that welfare in the abstract is an altogether inappropriate object of concern, and you therefore have *no* reason whatsoever to care about it.[19] I think this is too strong. I think you have some reason to care about welfare as such.[20] However, if my account above is right, acting out of such higher-order concern is deeply inauthentic. It takes the problem of inauthenticity—of orienting your actions around a value whose grounds you fail to appreciate—to the extreme. It's a bit as if you celebrated the value of Parfit's *Reasons and Persons* without even knowing whether it's a book of philosophy, a painting, or a pretentiously misnamed orphanage. It's a bit as if Leah played her well-memorized Zen-role without the faintest hunch of what it could all be about, just following the testimony of a reliable friend who told her such a life was good.[21]

[19] For a related skeptical view regarding *de dicto* concern about rightness, see e.g. Smith (1994) or Zhang (2021). This view is often motivated by a 'buck-passing account' of rightness (see e.g. Dancy 2000; Stratton-Lake 2002: 15). The skeptical view about welfare might be motivated by a parallel account of welfare (see e.g. Sidgwick 1907: 112; Darwall 2002; Skorupski 2007; or Fletcher 2012).

[20] For a related permissive view regarding *de dicto* concern about moral rightness, see e.g. Copp (1997), Lillehammer (1997), Svavarsdóttir (1999), Carbonell (2013), or Johnson King (2020).

[21] Thus, I think the intuitive problem of fetishism (as suggested e.g. by Smith 1994 or Zhang 2021) is one of authenticity. I hope to explore this idea in more depth in the future. For related ideas, see e.g. Hopkins (2007), Hills (2009), Mogensen (2017), Johnson King (2019; 2020), or Heering (2022). One might think there's an intermediate level. Perhaps we can't care authentically about the highest-order property of people being well off, or about fully concrete instances of schmusic, merging, or post-humaneness. But, one might think, these things will be instances of familiar *categories* of welfare-goods—forms of art, relationships, or virtues—and we can care authentically about these intermediate general categories. But this seems dubious. Suppose you find Arnold Schönberg's music

Again, promoting the welfare of contemporaries won't be nearly as inauthentic. True, we don't all share the exact same goals, desires, and values. But we contemporaries are still pretty similar, and can largely appreciate each other's good. So, if the inauthenticity of an action mitigates your reasons for it, the second premise of the above argument is false: *ceteris paribus*, you have stronger reason to care about the welfare of our contemporaries than about that of far-future people. To what extent this undermines the longtermist conclusion is a more complicated question. I'll get back to it in section 5. But first, let me introduce a second reason to care less about far-future people.

4 Meaning

Suppose Egan donates his semen to a sperm bank in Buenos Aires. Djamila receives it, and raises the resulting child, Imani, alone. With enormous dedication, mindful decisions, and loving affection, she teaches Imani gratefulness and environmentalist virtues, supports her passion for the accordion, and helps her through crises in her friendships and studies. So Imani lives a very good life. At the age of 50, she wants to know her biological father, determines the identity and whereabouts of Egan, and meets him. He is delighted to see her. He says he's always thought the meaning of life was to make the world better. And fathering such a happy person contributes so much to that. Indeed, he proclaims in enthusiasm, the value of Imani's life contributes just as much meaning to his life as to Djamila's.

Intuitively, that last claim is wrong. Imani's welfare contributes more meaning to Djamila's life than to Egan's. Let's see why that might be so. Intuitively, Egan is right insofar as the meaningfulness of your life isn't *just* due to how meaningful you subjectively find it. If you dedicate your life to counting grass or killing foreigners it isn't meaningful, even if you find it so. Promoting objectively good things is an important part of life's meaning. But how exactly is it? One might adopt a simple consequentialist account, on which all actual good consequences of your actions contribute equally (or just in proportion to their objective goodness) to the meaning of your life.[22] This, we might suppose, is Egan's view. But it's overly simplistic. Imani's life was a consequence of both Egan's and Djamila's actions, but provides more meaning for Djamila. Alternatively, one might adopt an *ex ante*, expectationalist interpretation, on which it's the expected value of your actions at the time of acting that provides the meaning. One might suggest that Egan couldn't have expected to produce a life like Imani's, while Djamila could. But that too seems false. Perhaps Egan had ample data on children from that sperm bank, and could reasonably expect to produce a life like Imani's. Still, Imani seems to provide more meaning for Djamila. More promisingly perhaps, one might say Djamila made a greater causal contribution to Imani's life than Egan. She not only contributed genetic material, but was also responsible for the specific

unnerving, but zealously promote it—simply because you care about 'art' in general, or about Frida Kahlo's paintings and Sophocles's tragedies. That seems inauthentic. Plausibly, the value of 'art' in general (insofar as it has any value at all) is grounded in the value of specific artworks, rather than vice versa. So, if you appreciate specific artworks, you'll only have a very partial appreciation of the grounds of the value of art. And you can't appropriately care about some artwork just in virtue of its being art. Thus there'll be something inauthentic if you promote schmusic, merging, or post-humaneness *simply* in virtue of their being art, relationships, or virtues.

[22] For such a view, see e.g. Smuts (2016).

virtues, relationships, and projects that formed her daughter's life. And this difference in degrees of causal contribution, one might say, is why Djamila derives more meaning. But this still seems too crude. Suppose Egan influenced Imani's life deeply, but in an entirely fortuitous way. He once hailed a Taxi, which otherwise would have run over Imani further down the road. He caused the pub quarrel in which the parents of Imani's future husband first met. He once fired an employee, who then started a business and gave Imani the job of her life. Even if Egan thus affected Imani's life deeply, Djamila derives more meaning from it.

Intuitively, the problem in this last scenario is that Egan's contributions were too accidental. To capture this, I propose the

> *Manifestation account of meaning:* One's life is meaningful (at least in part) to the extent that it promotes valuable things, such that the value of those things, or its grounds, are manifestations of one as a person.

At least as a rough approximation, we might say that *x* is a manifestation of *y* to the extent that it's an actualization of *y*'s dispositions. Suppose you drop your friends' vase, it shatters, and while fetching the flinders from under the sofa your friend then finds a long-lost bracelet. The shattering of that vase manifests its fragility: it's an actualization of it. The discovery of that bracelet doesn't manifest the fragility of the vase: it's not an actualization of that disposition, but an accidental consequence of it. So, if something is a manifestation of you as a person, it's not merely an accidental causal effect that you've had. It's an actualization—and thus an expression, emanation, or reflection—of your values, talents, or character.

To illustrate this idea, take Imani's life. If an objective list theory is correct, Imani's life is good in virtue of her specific virtues, relationships, or projects. The fact that Imani had those things is very much a manifestation of Djamila as a person: an actualization of Djamila's dedication and mindful attention, her concern for friendships, music, environmentalist virtues, and so on. Djamila is like the sculptor shaping the clay of her daughter's life. Perhaps to some degree, the fact that Imani had a good life is a manifestation of Egan too: an actualization of his concern for producing happy people, say. But the grounds of Imani's welfare weren't emanations from his person to anything like the degree to which they were for Djamila. It wasn't due to him that Imani was so environmentally conscious, played the accordion with such dedication, or maintained her friendships instead of ending them. The extent to which Imani's life is a manifestation of Egan—to which he can deem himself responsible for her welfare in the relevant sense—is much smaller. He's like the patron *funding* the sculptor shaping the clay of her life. And that's true even if objectively, or in expected value terms, or even in degrees of causal influence, his contribution was the same.

The manifestation account thus says that Imani's existence contributes less meaning to Egan's life than to Djamila's. More generally, it says that being non-accidentally responsible for good things contributes to the meaning of life. As I understand the account, it also implies the reverse. Being non-accidentally responsible for bad things, in this form, detracts importantly from meaning. If Imani lived a miserable life, as a direct manifestation of her mother's carelessness or incompetence, that reduces the meaningfulness of Djamila's existence—and much more so than it diminishes the meaningfulness of Egan's life. Now plausibly, meaning in life is not all about such manifestation. It's also about subjectively

enjoying one's promoting good, or about overcoming challenges, or passively appreciating existing value.[23] So the manifestation account aims to capture only part of the grounds of meaning. Still, non-accidentally promoting the good seems an important contributor to meaning.[24] And intuitively, you have reason to live a meaningful life. So if all of this is right, you have reason to promote good (or avoid promoting bad) things, such that their values are manifestations of you as a person. In particular, other things equal, you have more reason to promote good (or avoid promoting bad) things in such ways than through more accidental chains.

This should suffice to elucidate the idea for now.[25] If it's correct, it impacts longtermist actions. Suppose you invest in AI safety today, and through a fortunate chain of events thus contribute to some marvelous life in a million years. Then perhaps to some minimal degree, the value of that future life manifests your forethought. You invested in AI safety precisely to allow far-future people to exist. And perhaps it wasn't 100% fortuitous that you contributed to it. Still, *very* largely, the grounds of the value of this far-future life are accidental from your perspective. Again, that's part of our assumptions. And as with authenticity, the distance here cuts deep. For a start, any details about why or how this life will be good can't possibly manifest you as a person. Insofar as the constituents of far-future welfare are unknown unknowns from your perspective, you can't have intended to promote them in particular. So these details can't emanate from your specific concern. Moreover, and more simply, there's just a gargantuan array of intermediate causal factors, which you can't control or foresee, and which swamp your contribution almost completely. You're a tiny cog in the astronomical machine of the millennia grinding the effects of your actions. And that's true even if, in expected or objective value terms, your actions matter a lot. Your attempt to derive equal meaning from a far-future individual would be a *bit* like Egan's attempt to derive meaning from Imani. More aptly perhaps, it would be like saying about one of our Neanderthal ancestors that our present-day benefits from enjoying classical music or displaying environmentalist virtues are just as significant for the meaningfulness of their lives as were their children's benefits from the parental warmth they gave them.

Again, that's quite different with contemporaries. True, even if you save someone through an Oxfam donation, say, many aspects about it are accidental from your perspective. But

[23] For the first idea, see most prominently Wolf (2010). For something like the third idea, see e.g. Lovett and Riedener (2024a). Perhaps in particular, it can add meaning to your life if you appreciate the value of your own non-accidental works. Intuitively, say, there can be something tragic if your project reaches success only after your death—as with Gregor Mendel, Emily Dickinson, or Franz Kafka. This might add another problem of meaning for longtermists: they won't ever see the fruits of their labour.

[24] One might doubt this. In particular, one might think other factors better explain our intuitions about Egan and Djamila. One might say Djamila's raising of Imani seems more meaningful (in large part) because it involves more of a challenge or achievement, or takes up more time, or because she takes more subjective enjoyment from Imani's life. But I doubt that these factors add very much, if the underlying causal connection is sufficiently thin. Suppose Egan's sperm bank is hidden somewhere in the Argentinian wilderness and enormously challenging to reach; he thus spends a full 18 years to get to it; and once Imani is born, he witnesses every step she takes (perhaps through her social media presence) and enjoys it just as much as Djamila. Such a life still seems considerably less meaningful than Djamila's—perhaps even somewhat pitiable. The manifestation account thus seems to capture an important part of our initial intuitions.

[25] For a related view of meaning, see Brogaard and Smith (2005). In a deeper exploration of the account, more would need to be said especially about what it is for you to be more or less manifest in something. For such details, and some more general motivation, see Lovett and Riedener (2025; 2024a; 2024b). One point may be worth making here. As indicated, manifestation isn't simply a probabilistic notion. Even if Egan could have predicted with certainty how Djamila will raise Imani, Imani's life would be more a manifestation of her than of him. Still, probabilities are obviously relevant. If some state of affairs resulted from your actions, it's generally more a manifestation of you if you could have been certain that it will result than if it was highly unlikely to.

you know relatively well about the possible effects of your action, and can promote them comparatively directly. The degree to which you can understand this contemporary's life as a manifestation of your concern is much greater than with far-future people. Conversely, if you do nothing and they die, the degree to which you should understand this as a manifestation of you is much greater than with the death or non-existence of far-future individuals. So, if the meaningfulness of an action provides a reason for it, you again have stronger reasons to promote the welfare of contemporaries than that of far-future people.[26]

5 Objections and clarifications

Let's take stock. I've introduced two reasons to care less about far-future people than even about contemporaries you have no special positive relation to. Treating their welfare in just the same way would be problematically inauthentic and meaning-sapping. This suggests that, *ceteris paribus*, we should give more weight to the interests of our contemporaries than to those of far-future people. My two assumptions from section 2 were relevant for both points. Suppose that, instead of an objective list theory, hedonism was correct. Then we could care authentically about far-future pleasure. And since our contributions to far-future pleasure might not be quite as contingent as our contributions specifically to schmusic, the schmetaverse, or post-humaneness, we might derive a little more meaning from promoting *that*. Similarly, suppose we could reasonably assume that future lives will be similar to ours, and could have a more direct influence on them. Well, then we could authentically care about future friendships and virtues and so on. And our promoting them wouldn't be as accidental. So suppose my assumptions are true and my arguments sound.[27] What exactly does that mean for the longtermist conclusion? One might worry it doesn't mean much. So let me now clarify the nature, scope, and limitations of my arguments by answering three salient worries.

First, one might doubt that our two considerations should have any normative weight at all, or that I've even shown we should give *any* priority to our contemporaries. One might think concerns about authenticity and meaningfulness are simply self-indulgent. Ideally,

[26] These negative examples in particular suggest another possible argument against longtermism. One might take manifestation-relationships to underpin *moral responsibility*. For instance, one might say you're morally responsible for good or bad things to the extent that the grounds of their goodness or badness are manifestations of you as a person. So one might argue that we're generally more morally responsible for bad things in the present than for such things in the very far future, and that we thus have stronger reasons to prevent the former. This line of reasoning is perhaps more controversial, but I think it's promising, and I intend to explore it in future work.

[27] There's a worry that comes from questioning my second assumption. One might say it isn't *certain* that far-future lives will be radically different. And suppose there's even just a 1 in 10,000 chance that, conditional on our survival, our descendants will live largely as we do. Then longtermist actions might expectably still affect 100 billion (10^{11}) people like us. And that, one might say, still trumps almost everything. So doesn't the longtermist argument hold if we focus on these conservative scenarios only? In reply, first, this would constitute an interesting shift in focus. As indicated, longtermists have typically emphasized how different the far future could be. If we rely on conservative scenarios alone, this is an important dialectical reorientation. Second, it's not clear that the argument will actually work—i.e., that the relevant probabilities will be high enough—if we just rely on these cases. Even relatively small changes in our way of life, our capacities or values, can impact the nature of our welfare deeply. And our status quo is marked, on the one hand, by countless imperfections and problems (unlike perhaps the state of gods) and on the other hand, by ample knowledge and technology to address them (unlike perhaps the state of non-human animals or our distant ancestors). It just seems *very* unlikely that our form of life will stay largely intact for hundreds of thousands of years. Third, this worry anyway doesn't seem to affect the second point about meaning. Even if we do survive billions of years largely unchanged, that would still be almost entirely due to forces outside of our present actions.

you should care about others: you should try to do them as much good as you can, regardless of how you personally appreciate the way they live, or how your efforts contribute to *your* life. You should not—not even *ceteris paribus*—give any weight to such egocentric concerns.[28]

Let me say two things in reply. First, even granting that our considerations are largely self-centered, at a minimum, they seem to ground prerogatives. They set limits to what morality can demand of us. In particular, among other things, there's something *alienating* about actions that lack such authenticity and meaning. If you act in response to a value you fail to appreciate, there's a distance between your outward action and your deeper concerns. It's like a lack of 'integrity' à la Williams (1973; 1981b): a form of ignoring what actually moves you, discounting your attachments, dissolving your own person under the dictates of the impersonal good. Similarly, if you promote some good in highly accidental manners, there's a distance between you and that good. It's reminiscent of *Entfremdung* à la Marx: a way of doing things without intrinsic or immediate meaning, of reducing your own agency to a cog in a vast causal machine, without relevant responsibility or connection to the ultimate product.[29] Intuitively, morality can't demand of us to ignore these concerns entirely. So at least sometimes, it seems permissible to give *some* priority to our contemporaries.

Second, and less concessively, however, it just needn't be self-centered to live in accordance with our two considerations. More precisely, it needn't be a concern for yourself which makes you act in authentic and meaningful ways. It can simply be a concern for the values you care about. And for this reason (if not for others too[30]) it needn't be a permissible defect, but can be a realization of the ethical ideal. To see what I mean, suppose Kaya is a rising politician from Eritrea. Having experienced how a lack of education can hamper, she committed herself to improving the country's schools. And she's got an exceptional talent for it: courage, dedication, charisma, and plenty of creative and effective ideas. Now she thinks she could do *slightly* more good by going into consulting and donating ample money to improve conditions on Chinese factory farms. (Let's say other salient options don't occur to her for now.) But she has no relationship whatsoever to China, let alone to its animals, and feels rather appalled by the world of consulting. Kaya may then decide for a career as a politician, not out of concern for *herself*, but simply out of commitment to her fellow Eritreans, or out of concern for the value of literacy or female education. This will orient her life around values she appreciates, and make it more manifest in the good she effects. And intuitively (at least if the difference in promoted good isn't too vast), such a choice may well be admirable: it may seem disloyal towards her co-citizens, or noncommittal about her own values, if she applied at Boston Consulting. Thus, considerations of authenticity and meaning don't just ground prerogatives for permissible but suboptimal egotism. They're

[28] The worry here is similar to the worry that nonconsequentialist morality is 'self-indulgent'. For discussions of this worry, see e.g. Williams (1981a) or Blustein (1991).

[29] At one point, Marx describes the opposite of non-alienated production in words strikingly resembling the manifestation account: 'Let us suppose that we had carried out production as human beings. [. . .] In my *production* I would have objectified my *individuality, its specific character*, and therefore enjoyed not only an individual *manifestation of my life* during the activity, but also when looking at the object I would have the individual pleasure of knowing my personality to be *objective*' (Marx and Engels 1975: 227; emphasis in original).

[30] A simpler reason for why it can be admirable to live in accordance with our considerations might be that some amount of self-centeredness is just part of the well-lived life. For instance, perhaps it can manifest a dubious lack of self-respect to ignore our considerations entirely.

not categorically less worthy than concerns of impartial beneficence. They're a genuine part of the well-lived life.

But there's an obvious subsequent worry. Perhaps it can be permissible, or even admirable, to accord with our principles when you could otherwise do *slightly* more good. But the longtermist rationale suggests that, in many situations, you could do *vastly* more good by taking some longtermist action—save trillions more people in expectation. And surely, one might say, considerations of authenticity and meaning can't have enough weight to outbalance *that*. More concretely, perhaps you should give the interests of contemporaries n times more weight than those of far-future people, for some n greater than 1. But for any reasonable such factor, the longtermist argument will still go through. Even if we count far-future people 100,000 times less, say, in light of their astronomical number they should still dominate our concern. In sum, our considerations can't really threaten the longtermist conclusion—or so one might object.

In response, it's worth noting first that such a factor-interpretation doesn't capture our initial intuitions. For any such factor, the longtermist argument stands, provided our actions affect sufficiently many future people. So on this reading, the problem with the argument in section 1 was, at most, that the septillion (10^{24}) of future people I imagined were still too few. But that surely wasn't the intuitive problem. The argument wouldn't have been any more commonsensical if I had spoken of an octillion (10^{27}) or a decillion (10^{33}) to start with. Intuitively, something stronger is true. *No matter* how many far-future people our actions expectably affect, our contemporaries aren't generally practically irrelevant. In other words, at least sometimes, a certain form of non-aggregationism or lexical priority applies.[31]

That's our initial intuitions. But is it plausible that we should give lexical priority to our contemporaries, in light of the above concerns? I very much doubt such priority is always warranted, or that far-future people are generally practically irrelevant. I've suggested that the well-lived life isn't simply or primarily about impartial beneficence. But surely it isn't simply or primarily about authenticity or meaning either. Plausibly, the relative weight of different considerations must ultimately be determined in the context of entire lives. Currently, most agents *never* take actions with explicitly longtermist goals. And it's not plausible that our lives overall become problematically inauthentic and meaningless if we take some—or indeed quite considerable—longtermist action. I thus think our considerations of authenticity and meaning are absolutely compatible with the plausible view that we should spend *much* more resources on the long-term future than we currently do, and partly in light of the sheer number of possible people.[32]

However, I do think they cast doubt on the more radical idea that our contemporaries are *generally* practically irrelevant, or relevant only as tie-breakers. Suppose we literally always make our choices depend on longtermist considerations, and follow shorter-term goals only when these considerations are exactly tied. *Then* our lives would be devoid of authenticity and meaning to an extent that seems problematic. Consider Kaya again, and suppose she realizes that various longtermist actions are expectably *much* better than

[31] For related kinds of non-aggregationism, see e.g. Scanlon (1998: 235) or Voorhoeve (2014).

[32] As indicated in fn. 3, by 'longtermism', some people mean the view that the long-term future should be *a* key moral priority of our time. Depending on what it means to be 'a key priority', I think my arguments are thus compatible with this view.

improving either education in Eritrea or animal welfare in China. So she tries to follow the longtermist rationale as well as she can: in deciding which job to take, what to do with her money and spare time, whether to prioritize the climate crisis in her action or present-day racism or education in Eritrea, which politicians to elect or policies to support, whether or not to engage in comparatively minor lies or acts of defraud or betrayals of trust, and so on. She often does things with little short-term benefits at all. And when her actions do yield short-term goods, she regards them as merely instrumentally valuable, or as insignificant tie-breakers. *Then*, it seems, the grand orientation of Kaya's life would be worryingly alienated from anything whose goodness or badness she can genuinely value or have any direct influence on. At a minimum, that's not a life that can be demanded of her. More strongly, it's not the life she has most reason to live. And that's true regardless of how many far-future people are at stake. At some point, in some contexts, a lexical priority seems right.[33]

When exactly that's so—between total neglect and total prioritization of longtermist concerns—is naturally difficult to pinpoint. Perhaps we're permitted, and often have most reason, to live our lives *by and large* in response to goods we can appreciate relatively well and affect relatively directly. Or conversely, we're permitted, and often have most reason, to live our lives *by and large* without causing or allowing bad things we can appreciate rather well and affect rather directly. In any case, I think these considerations of authenticity and meaning shape our normative landscape importantly. For all the longtermist action we should plausibly undertake, we often have most overall reason to be guided by nearer-term concerns. One point may need emphasis at this juncture. My argument isn't simply a demandingness worry against strong altruistic obligations. It's compatible with very wide-ranging duties of beneficence. Specifically, suppose we could somehow rather directly save quadrillions of contemporaries. Plausibly, this would put extremely wide-ranging demands on us: it would outweigh almost any competing moral or non-moral concerns. My argument is perfectly compatible with this. It only implies that such demands aren't equally (or nearly) as weighty if these 'people' are profoundly different and distant.

Still, there's a third worry. *My* conclusions might just seem deeply counter-intuitive. I've claimed that a life oriented too much around longtermist concerns must be inauthentic and meaning-deprived. But this, one might say, just seems false. Quite to the contrary, it seems that *few* things could be of more genuine concern to us than our long-term trajectory. Few

[33] One might counter that longtermism wouldn't be as radically revisionist and alienating in practice. One might say we're almost totally clueless about the long-term effects of our actions. This means that in most situations, all of our options are expectably equally good in the longer term, or that the expected longer-term stakes (or value-differences among options) are generally small. So, one might argue, in most situations we can use near-term effects as tie-breakers, live largely as we live now, and thus avoid any serious alienation. In response, first, this isn't the view that longtermists typically take. They typically emphasize that their view is highly revisionist in practice (see e.g. Greaves and MacAskill 2021; MacAskill 2022: esp. ch. 10). Second, that longtermism is highly revisionist just seems very plausible, once you accept that *some* actions are expectably exceptionally good in the longer term. Suppose AI safety investments are generally exceptionally good. Well, you always have the option of trying to earn as much money as you can, and donate it for AI safety. And thus the long-term stakes almost always seem very high. Third, plausibly, even if it implied that our outward lives could remain largely as they are, longtermism would still have highly revisionist implications about the *attitudes* we should have towards our lives. It would imply we should see all of our standard concerns as insignificant tie-breakers—as *almost* totally and absurdly trivial in comparison to what really matters (see e.g. Nagel 1971). Plausibly, there'd still be something alienating about this. We cannot authentically see our own actions as insignificant in this sense. And doing so seems to threaten our sense of meaning. I'd love to explore these implications of longtermism for our attitudes on another occasion.

things could be as meaningful as trying to safeguard our long-term future. Perhaps Kaya currently cares a lot about the impact of menstruation on girl's absenteeism in the schools of Asmara, and derives meaning from addressing that. And perhaps we all focus on such parochial issues now. But couldn't we *possibly* come to care just as—or indeed much more—authentically and meaningfully about the long-term future of life in the universe? Taking a step back from my arguments, it may seem absurd to deny this.[34]

I think there's something to this intuition. But it doesn't vindicate the longtermist conclusion, or undermine my points. I've addressed one argument for longtermism only: the argument that grounds the importance of longtermist actions on the sheer number of far-future people. And here I stick to my claims. We just cannot care as authentically about far-future individuals, or derive as much meaning from doing them good, as with contemporaries. If we think we can, we must be underestimating their difference or distance from us, or employing some incorrect theories of authenticity and meaning. Now importantly, there could be *other* arguments for caring about the long-term future, which don't hinge on total future welfare. To take just one example, consider the long-term flourishing of humanity as a species. Perhaps the fulfillment of our species-potential has a value in itself, which doesn't even partly reduce to the value of our total individual welfare.[35] And so perhaps our long-term species-flourishing can be a goal whose value we can genuinely appreciate, and in whose realization we can be manifest rather distinctly. All of this is compatible with my argument. However, it seems implausible that such alternative rationales can make longtermist actions nearly as important as the standard argument *prima facie* did. Our long-term species-flourishing may be good. But it doesn't seem *so* valuable as to dominate all nearer-term concerns. Why should it be? What made the long-term future seem *so* important, on the standard argument, was simply that it contained so many people. To the extent that our concern for the far future becomes independent of the number of future individuals, it becomes less overwhelmingly significant.

6 Conclusion

Let me conclude. I've focused on one assumption of the standard longtermist argument: the idea that we should care as much about far-future people as about our contemporaries. I've argued against this premise: considerations of authenticity and meaning suggest we generally ought to care less about them. And I've claimed that this means, at least, that the far future shouldn't be as dominating a concern in practice as the longtermist argument initially suggests. We should plausibly take the far future much more seriously than we currently do. But at the same time, in many situations, even if longtermist actions would expectably do more good, we ought to act for contemporaries' sakes. Consider the United States again. Plausibly, the US should ramp up investments in AI safety. At the same time, it seems right for them to make it one of their priorities now to help Ukraine, simply because of the Ukrainians, and even if it isn't optimific in the very long term.

[34] I thank Jacob Barrett for helping me see this.
[35] For similar thoughts, see e.g. Rolston (1985), Frick (2017), or Ord (2020: 52–53).

Acknowledgments

For very helpful inputs or comments on this material, I thank Sarah Arrenbrecht, Jacob Barrett, Samuel Hughes, Harvey Lederman, Harry Lloyd, Adam Lovett, Andreas Mogensen, Lukas Naegeli, Peter Schaber, Christian Tarsney, and especially Hilary Greaves. I also thank the Global Priorities Institute for hosting me and providing a very inspiring environment while I was working on this chapter.

References

Archer, A. (2017), 'Integrity and the Value of an Integrated Self', in *Journal of Value Inquiry* 51/3: 435–454.
Bauer, K. (2017), To Be or Not to Be Authentic. In Defence of Authenticity as an Ethical Ideal', in *Ethical Theory and Moral Practice* 20/3: 567–580.
Blustein, J. (1991), 'Integrity and Self-Indulgence', in J. Blustein, *Care and Commitment: Taking the Personal Point of View* (Oxford University Press), 82–89.
Bostrom, N. (2008), 'Letter From Utopia', in *Law and Ethics of Human Rights* 2/1: 1–7.
Brogaard, B. and Smith, B. (2005), 'On Luck, Responsibility and the Meaning of Life', in *Philosophical Papers* 34/3: 443–458.
Broome, J. (1991), *Weighing Goods* (Blackwell).
Cancian, M. (2022), 'Aid to Ukraine Explained in Six Charts', *Center for Strategic and International Studies*, https://www.csis.org/analysis/aid-ukraine-explained-six-charts (accessed 11 January 2023).
Carbonell, V. (2013), 'De Dicto Desires and Morality as Fetish', in *Philosophical Studies* 163/2: 459–477.
Copp, D. (1997), 'Belief, Reason, and Motivation: Michael Smith's 'The Moral Problem', in *Ethics* 108/1: 33–54.
Dancy, J. (2000), 'Should We Pass the Buck?', in *Royal Institute of Philosophy Supplement* 47: 159–173.
Darwall, S. (2002), *Welfare and Rational Care* (Princeton University Press).
De-Shalit, A. (1994), *Why Posterity Matters: Environmental Policies and Future Generations* (Routledge).
Finnis, J. (1980), *Natural Law and Natural Rights* (Oxford University Press).
Fletcher, G. (2012), 'Resisting Buck-Passing Accounts of Prudential Value', in *Philosophical Studies* 157/1: 77–91.
Fletcher, G. (2013), 'A Fresh Start for the Objective-List Theory of Well-Being', in *Utilitas* 25/2: 206–220.
Frick, J. (2017), 'On the Survival of Humanity', in *Canadian Journal of Philosophy* 47/2–3: 344–367.
Greaves, H. (2017), 'Discounting for Public Policy: A Survey' in *Economics and Philosophy* 33/3: 391–439.
Greaves, H. and MacAskill, W. (2021), 'The Case for Strong Longtermism', GPI Working Paper No. 5-2021 (Global Priorities Institute, Oxford University).
Greaves, H., MacAskill, W., and Thornley, E. (2021), 'The Moral Case for Longterm Thinking', in N. Cargill and T. M. John (eds.), *The Long View: Essays on Policy, Philanthropy, and the Long-Term Future* (FIRST), 19–28.
Gregory, A. (2015), 'Hedonism', in G. Fletcher (ed.), *The Routledge Handbook of Philosophy of Wellbeing* (Routledge), 113–123.
Heathwood, C. (2015), 'Desire-Fulfillment Theory', in G. Fletcher (ed.), *The Routledge Handbook of Philosophy of Well-Being* (Routledge), 135–147.
Heering, D. (2024), 'Why and When is Pure Moral Motivation Defective', in *Erkenntnis* 89/2: 665–684.
Hills, A. (2009), 'Moral Testimony and Moral Epistemology', in *Ethics* 120/1: 94–127.
Hopkins, R. (2007), 'What Is Wrong With Moral Testimony?', in *Philosophy and Phenomenological Research* 74/3: 611–634.
Jacobs, A. (2018), *Thanks a Thousand* (TED Books).
Johnson King, Z. (2019), 'We Can Have Our Buck and Pass It, Too' in *Oxford Studies in Metaethics* 14: 167–188.
Johnson King, Z. (2020), 'Praiseworthy Motivations', in *Noûs* 54/2: 408–430.
Lillehammer, H. (1997), 'Smith on Moral Fetishism', in *Analysis* 57/3: 187–195.
Lloyd, H. (2021), 'Time Discounting, Consistency, and Special Obligations: A Defence of Robust Temporalism', GPI Working Paper No. 11-2021 (Global Priorities Institute, Oxford University).
Lovett, A. and Riedener, S. (2024a), 'The Good Life as the Life in Touch With the Good', in *Philosophical Studies* 181/5: 1141–1165.

Lovett, A. and Riedener, S. (2024b), 'Commonsense Morality and Contact with Value', in *Philosophy and Phenomenological Research* 109/1: 410–430.
Lovett, A. and Riedener, S. (2025), 'Touching the Good: Special Relationships as Contact with Value', *Journal of Ethics and Social Philosophy* 30/2: 179–204.
MacAskill, W. (2018), 'What Are the Most Important Moral Problems of Our Time?', *TED*, https://www.youtube.com/watch?v=WyprXhvGVYk (accessed 11 January 2022).
MacAskill, W. (2022), *What We Owe the Future* (Basic Books).
Marx, K. and Engels, F. (1975), *Collected Works, Vol. 3, 1843–44* (Intl Pub Co).
Mogensen, A. L. (2017), 'Moral Testimony Pessimism and the Uncertain Value of Authenticity', in *Philosophy and Phenomenological Research* 95/2: 261–284.
Mogensen, A. L. (2022), 'The Only Ethical Argument for Positive δ? Partiality and Pure Time Preference', in *Philosophical Studies* 179/9: 2731–2750.
Nagel, T. (1971), 'The Absurd', in *Journal of Philosophy* 68/20: 716–727.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Parfit, D. (1984), *Reasons and Persons* (Oxford University Press).
Paul, L. A. (2014), *Transformative Experience* (Oxford University Press).
Pogge, T. (2002), *World Poverty and Human Rights* (Polity Press).
Ramsey, F. P. (1928), 'A Mathematical Theory of Saving', in *Economic Journal* 38/3: 543–555.
Rawls, J. (1999), *A Theory of Justice* (Belknap Press).
Rolston, H. (1985), 'Duties to Endangered Species', in *BioScience* 35/11: 718–726.
Sanger, D. E., Troianovske, A., Barnes, J. E., and Schmitt, E. (2022), 'Ukraine Wants the U.S. to Send More Powerful Weapons. Biden Is Not So Sure', in the *New York Times*, https://www.nytimes.com/2022/09/17/us/politics/ukraine-biden-weapons.html (accessed 11 January 2023).
Scanlon, T. M. (1998), *What We Owe to Each Other* (Harvard University Press).
Sidgwick, H. (1907), *The Methods of Ethics* (Macmillan).
Smith, M. (1994), *The Moral Problem* (Blackwell).
Smuts, A. (2016), *Welfare, Meaning, and Worth* (Routledge).
Skorupski, J. (2007), 'Buck-Passing About Goodness', in T. Rønnow-Rasmussen, B. Petersson, J. Josefsson, and D. Egonsson (eds.), *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz* (Department of Philosophy, Lund University), 1–15.
Stratton-Lake, P. (2002), 'Introduction', in P. Stratton-Lake (ed.), *Ethical Intuitionism: Re-Evaluations* (Oxford University Press), 1–28.
Svavarsdóttir, S. (1999), 'Moral Cognitivism and Motivation', in *The Philosophical Review* 108/2: 161–219.
Taylor, C. (1992), *The Ethics of Authenticity* (Harvard University Press).
Varga, S. (2011), *Authenticity as an Ethical Ideal* (Routledge).
Voorhoeve, A. (2014), 'How Should We Aggregate Competing Claims?', in *Ethics* 125/1: 64–87.
Williams, B. (1973), 'A Critique of Utilitarianism', in J.J.C. Smart and B. Williams, *Utilitarianism: For and Against* (Cambridge University Press), 77–150.
Williams, B. (1981a), 'Utilitarianism and Moral Self-Indulgence', in B. Williams, *Moral Luck* (Cambridge University Press), 40–53.
Williams, B. (1981b), 'Persons, Character and Morality', in B. Williams, *Moral Luck* (Cambridge University Press), 1–19.
Wolf, S. (2010), *Meaning in Life and Why It Matters* (Princeton University Press).
Woodard, C. (2015), 'Hybrid Theories', in G. Fletcher (ed.), *The Routledge Handbook of Philosophy of Well-Being* (Routledge), 161–174.
Zhang, X. (2021), 'Why De Dicto Desires Are Fetishistic', in *Ratio* 34/4: 303–311.

PART 2

PREDICTING AND EVALUATING THE FUTURE

10

What Are the Prospects of Forecasting the Far Future?

David Rhys Bernard and Eva Vivalt

1 Introduction

There has been a recent increase in interest in longtermism.[1] Longtermism postulates that if one is trying to maximize well-being, one of the most important things to consider is the impact of one's actions on the far future, due to the potential scale of the far future. However, in order to be able to positively affect the far future, we need to be able to at least partially predict the impacts of our actions over this time period. In other words, future generations may increase the stakes of our actions exponentially, but if our accuracy about the impacts of our actions also decreases exponentially over time, it is not clear which factor will win out.

Unfortunately, it is very difficult to know how good one's forecasts might be over a very long time frame. It is time-consuming and impractical to collect forecasts over such a time period and wait for them to resolve, and even if one did, forecast ability could itself change over time such that one's findings would not be externally valid. Forecast ability over long durations may also depend on the kinds of items being forecast.

Nonetheless, while it is clearly impossible to fully evaluate the merits of longtermism on an empirical basis, we believe it is worthwhile to consider whether there is a 'decay effect' in the accuracy of forecasts of impacts over the short run. Such an analysis would necessarily be limited, but if we observed that experts could not accurately forecast the impacts of actions over even a few years, it might update our beliefs about our ability to forecast impacts over a much longer time frame.[2]

We focus on forecasts of the results of impact evaluations—studies that seek to estimate the causal impact of some action on an outcome—rather than forecasts of other phenomena like election outcomes. This is because if we are interested in using forecasts to inform actions, our ability to forecast the impacts of actions will be more directly relevant than our ability to forecast what would happen in the absence of any particular action. In other words, we argue that we should distinguish between 'state' forecasts, such as what net carbon emissions will be in 2050, and 'causal' forecasts, such as what effect a certain policy will have on reducing net carbon emissions in 2050. As the name suggests, the latter refers to a *causal* relationship. We will also distinguish between 'causal' forecasts and 'conditional'

[1] E.g., Greaves and MacAskill (2021), Bajekal (2022), MacAskill (2022).

[2] It should be noted that even in the absence of the ability to make long-run forecasts, short-run forecasts could be useful. For example, Millner and Heyen (2021) argue that short-term forecasting ability can be more important than long-term forecasting if it substitutes for it in practice.

David Rhys Bernard and Eva Vivalt, *What Are the Prospects of Forecasting the Far Future?* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press.
© David Rhys Bernard and Eva Vivalt 2025. DOI: 10.1093/9780191979972.003.0010

forecasts. 'Conditional' forecasts are forecasts that ask about the state conditional on some other state or set of states.

2 State, conditional, and causal forecasts

To make the distinction clearer, we will leverage the framework of graphical models. We can represent a 'state' forecast such as 'What is the likelihood that Y = 1' as simply predicting a value for Y (Figure 10.1):

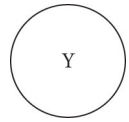


Figure 10.1 State forecasts.

We might nest our prediction of Y in some model of how the world works. For example, we might think that X_1 affects Y, so that the state of X_1 is important in predicting Y (Figure 10.2):

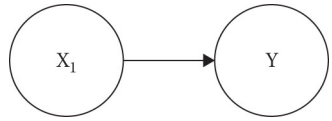


Figure 10.2 Conditional forecasts.

In this case, we might ask questions of the form 'What is the likelihood of Y = 1 if $X_1 = 1$?' We will refer to these kinds of forecasts as 'conditional' forecasts.

Even if we knew the conditional probability that Y = 1 given $X_1 = 1$ and the conditional probability that Y = 1 given $X_1 = 0$, it would not be enough to determine the impact that a change in the value of X_1 would have on the likelihood that Y = 1. For example, suppose we observed Y and X_1 at a given point in time and determined that the conditional probability of observing Y = 1 given $X_1 = 1$ was 0.9, or $E(Y = 1|X_1 = 1) = 0.9$. This would be a statement about correlations, not causality; some unobserved factor could be driving the correlation between X_1 and Y, such that increasing $X_1 = 0$ to $X_1 = 1$ would not affect Y by as much as it seems, or perhaps not at all. For example, perhaps the real model of the world looks like Figure 10.3:

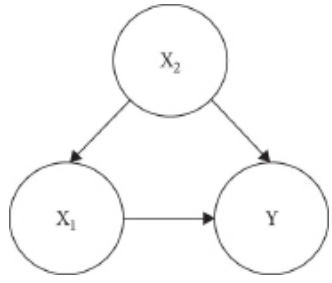


Figure 10.3 Sample true model

In this case, conditional forecasts could be misleading.

Causal forecasts, by contrast, can be thought of as isolating the causal impact of X_1 on Y. For example, someone might run an experiment which randomizes assignment of X_1 to observe its effect on Y. This would be represented by a graphical model of the following form (Figure 10.4):[3]

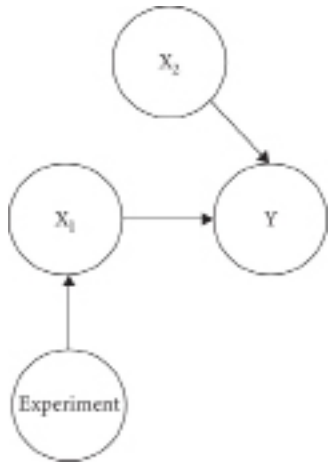


Figure 10.4 Causal forecasts.

While causal forecasts retain the limitation that they do not necessarily say what the impact of changing X_1 would be under different circumstances,[4] they are clearly close to what one would ideally want in order to inform decisions. This is not to say that state or conditional forecasts cannot be useful. For example, state or conditional forecasts of the likelihood of a volcanic eruption or nuclear war could improve preparedness. When we know the outcomes of our potential actions very well, perhaps because the outcomes mechanically follow from the action, causal forecasts may not be needed.

In summary, state, conditional, and causal forecasts are theoretically distinct, and one's ability to make accurate state or conditional forecasts is not necessarily indicative of one's ability to make accurate causal forecasts.[5] One could be very good at forecasting what would happen should current trends continue in the absence of intervention, but worse at making forecasts of what would happen if one intervened. Conversely, in some cases it may be easier to predict what would happen should one take an action than in the absence of intervention. The correlation between state, conditional, and causal forecast accuracy ultimately is an empirical question, on which there is limited evidence.[6]

[3] The interested reader is referred to Pearl (2009) for an introduction to graphical models. While this chapter assumes causal decision theory, it should also be noted that in philosophy causal decision theory is an area of active debate.

[4] For example, the true model could be much more complex than the graph in Figure 10.4, and results may not generalize.

[5] If one's ability to make accurate state or conditional forecasts were correlated with one's ability to make accurate causal forecasts, it would be another reason to study the former, particularly as there are many more opportunities to gather data on the performance of state or conditional forecasts than data on the performance of causal forecasts.

[6] Preliminary results in DellaVigna and Vivalt (2025) suggest some correlation.

3 Data and approach

With this background in place, what does the evidence say about the accuracy of causal forecasts over time? We gathered data on forecasts connected to impact evaluations in the social sciences from several sources.

First, we scanned the list of studies that were posted on the Social Science Prediction Platform. This is a platform that researchers can use to elicit *ex ante* forecasts of what their studies will find.[7] Of the 37 studies that had been posted on the platform at the time of data collection, nine had results available which could be attached to the causal forecasts to gauge their accuracy.[8] Many studies could not be included because they were still in progress and had yet to make results available.

We supplemented these data with data from other studies we were aware of that had collected their forecasts outside of the platform, generally before the platform existed. There were 38 of these studies, of which 20 contained at least one estimate of a causal impact and had results available.[9] The studies identified and included through either of these paths are listed at the back of this chapter.

Combining the forecasts from both approaches, we were left with only 29 studies with estimates of causal forecasts that have resolved at time of writing. Clearly, any conclusions we draw from these data points will be extremely tentative. However, we find it still worthwhile to analyze them for what they can say, while including sensitivity analyses to emphasize how much we do not know. Apart from the small sample size, these studies may also suffer from selection bias if those that resolve particularly quickly have effects that are also easier or harder to predict on average.

Given our interest in how the quality of causal forecasts decays over time, we also construct a measure of how much time passed until the final data were collected. In particular, we define the relevant length of time to be the time from the end of the intervention to the time that the final data were collected.

In some circumstances, it could make sense to instead consider the time from the collection of the forecasts to the time the final data were collected. However, not all papers report details about when they elicited forecasts, or they provide a fairly large range. More importantly, for most of the studies in our sample,[10]we have no reason to believe that any results were made available between the intervention and when the final data were collected, so we do not believe that respondents would have substantively updated on post-intervention data.[11]

[7] One of the authors is a co-Principal Investigator on this platform, with Stefano DellaVigna. For further information on the purpose and some use cases of the platform, see DellaVigna, Pope, and Vivalt (2019).

[8] We did not use the microdata on the forecasts provided on the Social Science Prediction Platform, but rather collected forecasts and results directly from the papers themselves. This was done to make the data more comparable to the data collected from studies that were not posted on the Social Science Prediction Platform, in order to boost the sample size of forecasts available for analysis.

[9] Since collecting forecasts of causal estimates is still relatively new, there are no norms around how forecasts should be reported in papers. For example, some authors reported the mean forecast, while others reported median forecasts; we used only the mean forecast in our estimates, as more researchers reported this statistic.

[10] Bloom et al. (2020) is an exception to this.

[11] Many research teams include information in their forecasting surveys from a baseline survey conducted around the beginning of the intervention in order to help the forecasters make better forecasts, and it is generally regarded as good practice to do so whenever possible (DellaVigna, Otis, and Vivalt 2020); further, many studies in our data may have added forecasts at the last minute before releasing results. This means that counting the time between the collection of the forecasts and the final results could be misleading, as forecasters are likely to be basing their predictions primarily on information from around the time of the intervention. Overall, our measure

Some papers contain multiple causal forecasts that can be converted to effect sizes. Since the studies vary in the number of forecasted outcomes, but few study more than one time period, we collapse the forecast accuracy measures to the study-time period level to mitigate the risk that a single study drives results.[12] For our measure of accuracy, we focus on the Mean Absolute Percentage Error (MAPE), calculated as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(Forecast_i - Result_i)}{Result_i}\right|$$

This measure is frequently used in the literature and has the nice property of being relatively easily interpretable. For example, a MAPE of 1 means that the absolute difference between the forecast and the result is on average as large as the absolute value of the result being predicted. Focusing on the MAPE also allows us to combine estimates from many papers, as it is a unitless figure. If we were to instead restrict attention to those outcomes that were reported in or could be converted to common units (such as standard deviations), we would be able to combine data from at most eight papers. However, the MAPE has several limitations. First, it approaches infinity as the result approaches 0. Second, it is asymmetric, as it is bounded to a 100% error on the low side but unbounded on the high side. Third, it is a post hoc measure and not a proper scoring rule; forecasters would not have been incentivized to provide the most accurate forecasts.[13] Nonetheless, given the data limitations, it is the best we can do at present, and we hope future work will expand on this approach when more data are available.

4 Results

Figure 10.5 shows the MAPE of these 29 studies over time, plotted on a log scale. Notably, while even the longest time periods considered are not very long, the average prediction error seems to increase somewhat over time, pointing to the existence of a decay effect. If we were willing to assume that the decay effect continued at the same rate over longer time periods—a heroic assumption—the average MAPE would reach 3.8 for forecasts over 10-year

of timing will be more relevant than the time between the collection of the forecasts and the final data collection if forecaster accuracy does not depend much on information that was revealed between the intervention and the final data (such as the performance of the economy or other publicly available information), but rather depends more on the conditions that existed in the pre-intervention period that were likely reported in the forecasting survey and therefore salient to forecasters as they made their forecasts.

[12] Of the 29 studies, 26 only collect forecasts for one post-intervention time period. Two of the studies collect forecasts for two time periods and one study collects forecasts for three time periods.

[13] Indeed, in our set of studies, forecasters were often not incentivized at all in monetary terms. While there is reason to believe that academics who respond to surveys are more motivated by reputational concerns and intrinsic motivation than monetary incentives (Christiansen et al. 2019), we cannot rule out that forecasters facing higher incentives would perform better. However, we do not have reason to expect that any differences in forecast quality in response to higher incentives would vary by time period. It would be reasonable to think that even if overall forecast accuracy increased with higher incentives, it should not change the sign of the possible decay of forecast quality over time.

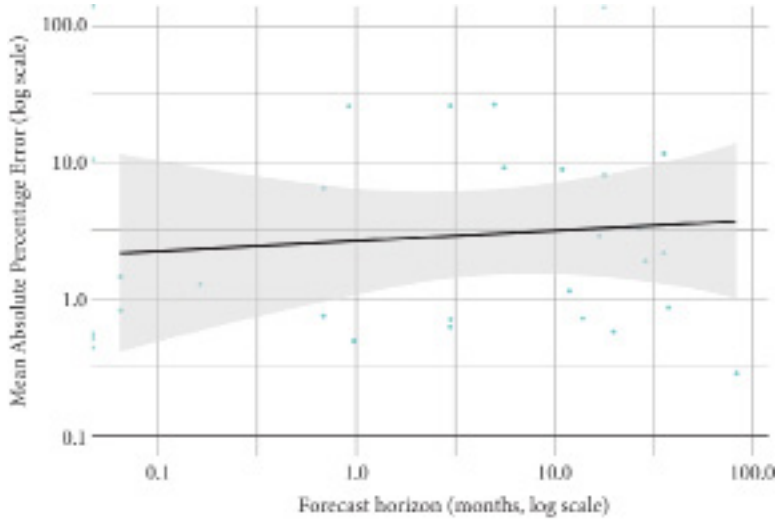


Figure 10.5 Mean Absolute Percentage Error.

periods and 4.6 for forecasts over 100-year periods, assuming a linear relationship between the log MAPE and log forecast horizon.

At the same time, we want to emphasize that these results are extremely tentative, given the small sample size of causal forecasts that have resolved. To demonstrate the uncertainty associated with this result, we perform a sensitivity analysis by randomly removing one study at a time from the data and examining the extent to which the coefficients of a regression of the MAPE on time vary. Figure 10.6 shows the results of this analysis. Depending on the study dropped, estimates of the coefficient on time range from 0.019 to 0.158.

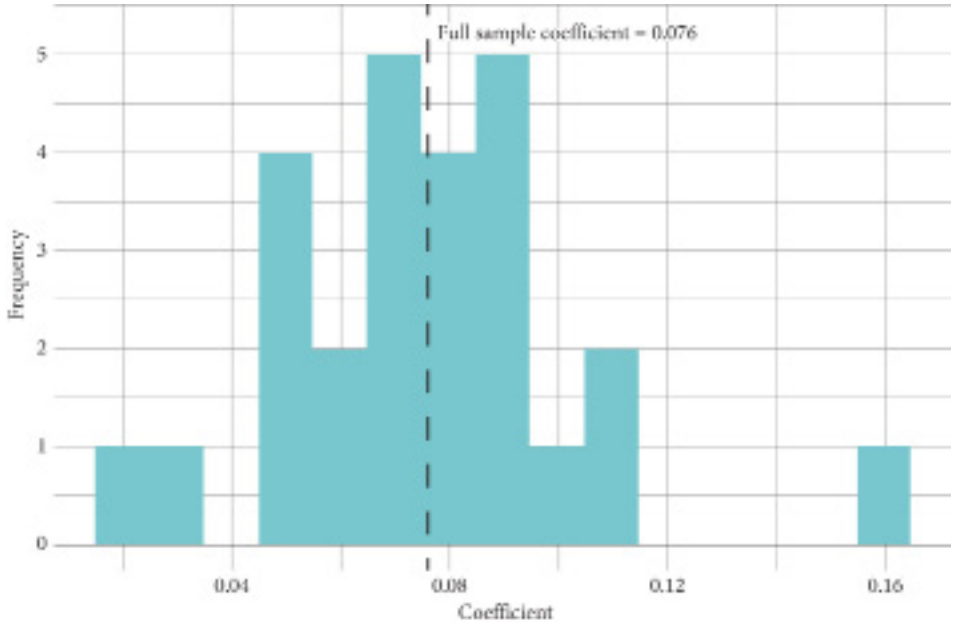


Figure 10.6 Sensitivity of coefficient to dropping one observation.

Apart from the forecasts being on a limited set of study results, we might expect forecast accuracy to change over time. For example, Chang et al. (2016), among others, have considered how to obtain more accurate forecasts, and it is possible that under optimal conditions forecasters could perform better. Nonetheless, to the extent to which our results reflect some inherent difficulty in forecasting the farther future, we would not expect the sign of the decay in forecast quality to flip under optimal conditions.

5 Discussion

What do our results imply for longtermism?

First, the most obvious takeaway is that there simply is not much data on the accuracy of causal forecasts, even in the short run. This would suggest the importance of epistemic humility with respect to longtermism: we don't really have much hard evidence, for or against, about forecast accuracy of causal effects over the long term.

Second, it should be noted that there are many different kinds of actions one can think of taking in pursuit of improving long-term outcomes. It is not clear that inaccuracy in making long-term forecasts undermines the case for mitigating all risks. For example, consider efforts to reduce the risk that an asteroid large enough to cause human extinction hits Earth. We may be reasonably confident that technology that could better detect and intercept an asteroid would remain helpful no matter what else happened between now and the period in which such a technology was needed.

There is some evidence that social programs have more 'stable' effects the more direct the path is between the intervention and the outcome (Vivalt 2020). In other words, if the true model of how the intervention is to have an effect involves the interaction of many variables, the variance of the results of different instantiations of the same program is likely to be higher. We may also expect individuals to be worse at forecasting the effects of such interventions. In plain English, 'more things can go wrong' with these interventions, so their effects should theoretically be harder to forecast, even in a single short-run forecast. In the longer term, we might think these forecasts decay more, too, as various factors that their impacts depend on may also be more liable to change.[14]

We can thus think of estimates of each intervention's effects as being associated with a latent level of *robustness to decay*. The interventions that we consider in our analysis are exclusively social science interventions whose main effects could be expected to depend a lot on context. Their effects might be thought of as generally having a lower level of robustness to decay than the effects of interventions with a more direct causal chain or a simpler graph.

As it stands today, the literature on forecasts provides little guidance regarding longtermism, one way or the other. Future work could explore the extent to which an increased focus on modeling could identify the situations in which forecasts are likely to be informative. Given the possibility that forecasts could decay in quality over time sufficiently quickly as to rule out some forms of longtermism, it is important to consider the robustness to decay of an estimate. Experimental work in the social sciences to date has largely ignored this issue.

[14] Intuitively, if these interventions' effects depend on more factors, then even holding constant the rate of change of each factor, we should expect their effects to vary more over time.

Bibliography

Abebe, G., Caria, A. S., and Ortiz-Ospina, E. (2021), 'The Selection of Talent: Experimental and Structural Evidence from Ethiopia', in *American Economic Review* 111/6: 1757–1806.

Allen, J., IV, Mahumane, A., Riddell, J., IV, Rosenblat, T., Yang, D., and Yu, H. (2021), 'Teaching and Incentives: Substitutes or Complements?' (National Bureau of Economic Research Working Paper Series).

Bajekal, N. (2022), 'Want to Do More Good? This Movement Might Have the Answer', in *TIME Magazine*. https://time.com/magazine/us/6206429/august-22nd-2022-vol-200-no-7-u-s/

Bedoya, G., Coville, A., Haushofer, J., Isaqzadeh, M., and Shapiro, J. P. (2019), 'No Household Left Behind: Afghanistan Targeting the Ultra Poor Impact Evaluation', WPS8877 (World Bank Working Paper Series).

Bessone, P., Rao, G., Schilbach, F., Schofield, H., and Toma, M. (2021), 'The Economic Consequences of Increasing Sleep among the Urban Poor', in *The Quarterly Journal of Economics* 136/3: 1887–1941.

Bloom, N., Mahajan, A., McKenzie, D., and Roberts, J. (2020), 'Do Management Interventions Last? Evidence from India', in *American Economic Journal: Applied Economics* 12/2: 198–219.

Campos-Mercade, P., Meier, A. N., Schneider, F. H., Meier, S., Pope, D., and Wengström, E. (2021), 'Monetary Incentives Increase COVID-19 Vaccinations', in *Science* 374/6569: 1–4.

Chang, W., Chen, E., Mellers, B., and Tetlock, P. (2016), 'Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments', in *Journal of Judgment and Decision Making* 11: 509–526.

Chopra, F., Haaland, I., and Roth, C. (2022), 'Do People Demand Fact-Checked News? Evidence from US Democrats', in *Journal of Public Economics* 205: 104549.

Christensen, G., Wang, Z., Paluck, E. L., Swanson, N., Birke, D., Miguel, E., and Littman, R. (2019), 'Open Science Practices are on the Rise: The State of Social Science (3S) Survey', working paper.

Christian, P. J., Garg, T., and Batmunkh, O. (2020), *RJK IE End Line Report: Final Impact Evaluation Report for Rani Jamara Kulariya Modernization Project - Nepal* (World Bank Group), https://documents.worldbank.org/en/publication/documents-reports/documentdetail/555881592475767388/rjk-ie-end-line-report-final-impact-evaluation-report-for-rani-jamara-kulariya-modernization-project-nepal. Accessed 9 January 2025.

de Andrade, G. H., Bruhn, M., and McKenzie, D. (2016), 'A Helping Hand or the Long Arm of the Law? Experimental Evidence on What Governments Can Do to Formalize Firms', in *The World Bank Economic Review* 30/1: 24–54.

Deb, B. C., Sircar, B. K., Sengupta, P. G., De, S. P., Mondal, S. K., Gupta, D. N., Saha, N. C., Ghosh, S., Mitra, U., and Pal, S. C. (1986), 'Studies on Interventions to Prevent Eltor Cholera Transmission in Urban Slums', in *Bulletin of the World Health Organization* 64/1: 127–131.

Del Carmen, G., Espinal Hernandez, E. E., and De Gouvea Scot De Arruda, T. (2022), 'Targeting in Tax Compliance Interventions: Experimental Evidence from Honduras', Policy Research Working Paper 9967 (World Bank Working Paper), https://openknowledge.worldbank.org/handle/10986/37155. Accessed 9 January 2025.

DellaVigna, S. and Pope, D. (2018), 'Predicting Experimental Results: Who Knows What?', in *Journal of Political Economy* 126/6: 2410–2456.

DellaVigna, S. and Vivalt, E. (2025). 'Forecasting Social Science: Evidence from 100 Projects', Working Paper.

DellaVigna, S., Otis, N., and Vivalt, E. (2020), 'Forecasting the Results of Experiments: Piloting an Elicitation Strategy' in *AEA Papers and Proceedings* 110/5: 75–79.

DellaVigna, S., Pope, D., and Vivalt, E. (2019), 'Predict Science to Improve Science', in *Science* 366/6464: 428–429.

Dimant, E., Clemente, E. G., Pieper, D., Dreber, A., Gelfand, M., and Behavioral Science Units Consortium. (2022), 'Politicizing Mask-Wearing: Predicting the Success of Behavioral Interventions among Republicans and Democrats in the US', in *Scientific Reports* 12/1: 7575.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. W. (2019), 'General Equilibrium Effects of Cash Transfers: Experimental Evidence from Kenya', Working Paper 26600 (National Bureau of Economic Research Working Paper Series), https://doi.org/10.3386/w26600

Greaves, H. and MacAskill, W. (2021), 'The Case for Strong Longtermism', Global Priorities Institute Working Paper 5-2021 (Global Priorities Institute, Oxford Unversity).

Groh, M., Krishnan, N., McKenzie, D., and Vishwanath, T. (2016), 'The Impact of Soft Skills Training on Female Youth Employment: Evidence from a Randomized Experiment in Jordan', in *IZA Journal of Labor & Development* 5/1.

Haushofer, J., Mudida, R., and Shapiro, J. P. (2020), 'The Comparative Impact of Cash Transfers and a Psychotherapy Program on Psychological and Economic Well-being', Working Paper 28106 (National Bureau of Economic Research Working Paper Series), https://doi.org/10.3386/w28106

Hirshleifer, S., McKenzie, D., Almeida, R., and Ridao-Cano, C. (2016), 'The Impact of Vocational Training for the Unemployed: Experimental Evidence from Turkey', *The Economic Journal* 126/597: 2115–2146.

Iacovone, L., Maloney, W., and McKenzie, D. (2022), 'Improving Management with Individual and Group-Based Consulting: Results from a Randomized Experiment in Colombia', in *The Review of Economic Studies* 89/1: 346–371.

Leong, K. C., Chen, W. S., Leong, K. W., Masturad, I., Mimie, O., Sheikh, M. A., Zailinawati, A. H., Ng, C. J., Phua, K. L., and Teng, C. L. (2006), 'The Use of Text Messaging to Improve Attendance in Primary Care: A Randomized Controlled Trial', in *Family Practice* 23/6: 699–705.

Liew, S.-M., Tong, S. F., Lee, V. K. M., Ng, C. J., Leong, K. C., and Teng, C. L. (2009), 'Text Messaging Reminders to Reduce Non-attendance in Chronic Disease Follow-up: A Clinical Trial', in *British Journal of General Practice* 59/569: 916–920.

Lua, P. L. and Neni, W. S. (2013), 'A Randomised Controlled Trial of an SMS-Based Mobile Epilepsy Education System', in *Journal of Telemedicine and Telecare* 19/1: 23–28.

MacAskill, W. (2022), 'The Case for Longtermism', in *New York Times*, 5 August 2022, https://www.nytimes.com/2022/08/05/opinion/the-case-for-longtermism.html

Manian, S. and Sheth, K. (2021), 'Follow My Lead: Assertive Cheap Talk and the Gender Gap', in *Management Science* 67/11: 6880–6896.

Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J. E., Brody, I., Chabris, C. F., Chang, E., Chapman, G. B., Dannals, J. E., Goldstein, N. J., Goren, A., Hershfield, H., Hirsch, A., et al. (2022), 'A 680,000-Person Megastudy of Nudges to Encourage Vaccination in Pharmacies', in *Proceedings of the National Academy of Sciences* 119/6.

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., Bonacorsi, L., Camerer, C., Chang, E., Chapman, G., Ciadini, R., Dai, H., Eskeris-Winkler, L., Fishbach, A., Gross, J. J., et al. (2021), 'Megastudies Improve the Impact of Applied Behavioural Science', in *Nature* 600/7889: 478–483.

Millner, A. and Heyen, D. (2021), 'Prediction: The Long and the Short of It', in *American Economic Journal: Microeconomics* 13/1: 374–398.

Orkin, K., Garlick, R., Mahmud, M., Sedlmayr, R., Haushofer, J., and Dercon, S. (2020), *Aspirations, Assets, and Anti-poverty Policies* (unpublished manuscript).

Pearl, J. (2009), *Causality* (Cambridge University Press).

Saccardo, S. and Serra-Garcia, M. (2020), 'Cognitive Flexibility or Moral Commitment? Evidence of Anticipated Belief Distortion', working paper, http://econweb.umd.edu/~davis/eventpapers./SerraGarciaFlexibility.pdf. Accessed 9 January 2025.

Samek, A. and Longfield, C. (2023), 'Do Thank-You Calls Increase Charitable Giving? Expert Forecasts and Field Experimental Evidence', in *American Economic Journal: Applied Economics* 15/2: 103–124.

Thomas, C. C., Otis, N. G., Abraham, J. R., Markus, H. R., and Walton, G. M. (2020), 'Toward a Science of Delivering Aid with Dignity: Experimental Evidence and Local Forecasts from Kenya', in *PNAS Proceedings of the National Academy of Sciences of the United States of America* 117/27: 15546–15553.

Vivalt, E. (2020), 'How Much Can We Generalize From Impact Evaluations?', in *Journal of the European Economics Association* 18/6: 3045–3089.

11

Taking the Long View

Paleobiological Perspectives on Longtermism

Rachell Powell

1 Two tyrannies

1.1 Evolved morality and the tyranny of the present

Human moral psychology is ill-equipped for the threats that humanity faces in the new millennium, from climate change and the collapse of biodiversity to thermonuclear war, synthetic bioweapons, and the rise of powerful artificial intelligence (AI). Human morality was evolutionarily designed to support cooperation within small, scattered hunter-gatherer groups comprised of moderately related individuals in an arena of discrete intergroup competition—radically different ecological circumstances than we now find ourselves in. The legacy of this adaptive history is a highly truncated capacity for other regard and a moral short-sightedness that prevent us from living up to the demands of our considered morality (Persson and Savulescu 2012). Although the development of social, political, and legal institutions since the Enlightenment has extended human altruism beyond the confines of kin, cooperative groups, and even relationships of reciprocity, humans remain highly susceptible to tribalistic moral reversions (Buchanan and Powell 2016; 2018). Monumental gains in moral inclusivity, such as the rule of law, democratic constitutionalism, political and civil liberties, and the ethical treatment of non-human animals, are all fragile and can be reversed on a dime, as recent sociopolitical trends painfully attest (Buchanan 2020; Kumar and Campbell 2022).

A truly inclusive morality goes well beyond the moral inclusion of large swaths of living beings, however. It also requires that we care about the effects of our present or near-term behavior on the lives of people that will come to exist in the future—individuals whom we will never meet, who will not bear a close genetic relation to us, who may not share our cultural values, and who by dint of the laws of physical causality will never be able to return the favor. Geography is no longer an excuse for moral inaction, given the ability to aid—and temporal distance should likewise be no excuse, given the ability to predict and intervene upon the future. Moral short-sightedness may not have mattered much for 2 million years of human evolution, during which time humans and their ancestors were limited in the causal influence they could exert on their surrounds. But it poses a grave ethical problem today, when human activities are altering the evolutionary course of life on Earth and threatening the very future of humanity. If we are to weather the spate of existential risks that lie before us (Ord 2020), we need the perspicacity as a species to forego the proverbial single marshmallow and to start securing value structures for the long haul.

Rachell Powell, *Taking the Long View* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Rachell Powell 2025. DOI: 10.1093/9780191979972.003.0011

This is the gist of a school of philosophy known as *longtermism* (MacAskill 2022). A strong version of the thesis, defended by Greaves and MacAskill (2021), maintains that the most important moral dimension of any proposed action or policy is the impact it will have on the long-term future, and further, that we have a moral duty to act on those impacts, since the loss in value from failing to safeguard the future of humanity is literally astronomical. We are living in a relatively early phase of the possible lifespan of our species, so the logic goes: thousands, perhaps millions of generations lie ahead of us, encompassing trillions or perhaps quadrillions of valuable lives. One's location in time is just as arbitrary a basis for discounting morally protectable interests as race, gender, religion, geography, and species membership. To strongly prioritize our own interests or those of our immediate descendants over the lives of helpless, voiceless, distant-future persons amounts to a 'tyranny of the present': an arbitrary disregard for the interests of an overwhelming majority of future people by a miniscule minority of nepotistic, short-sighted presently existing persons.

1.2 Considered morality and the tyranny of the future

The tyranny of the present poses such a grave threat to the future of human flourishing that some authors have proposed global biomedical moral enhancement (Persson and Savulescu 2012) and the rapid development of benevolent artificial superintelligence (Bostrom 2014) as last-ditch, Hail Mary attempts to safeguard the continued existence of the human species. The risks and potential rights violations associated with such proposals could be deemed justified on a wide range of moral views by the magnitude of moral loss that would be incurred if existential threats came to pass, no matter how infinitesimal or 'Pascalian' their probabilities might be (Thorstad 2021; Tarsney 2022).

This swamping effect, whether expressed quantitatively or qualitatively, becomes far more pronounced once we factor in the possibility of interstellar space travel. Humans have approximately 500 million years to enjoy planet Earth until the bloating sun begins to boil away the oceans, rendering the Earth unhabitable to complex multicellular life and ultimately consuming the planet. Yet even the sun's main sequence life cycle need not pose an outer boundary on humanity's future: for so long as we play our cards right, technological progress will march on unimpeded, and humanity will have taken to the stars, long before the end of the world. Spacefaring human civilizations could survive well beyond the death of the sun to take advantage of the multi-trillion-year span of star formation throughout the universe, allowing for a mind-boggling number of well-lived lives converting dead physical energy into living value structures over the vastness of cosmic space and time.

Even if these grand cosmic visions never materialize, longtermist arguments may still go through, since a sufficiently long future of humanity on Earth could be enough to sustain the swamping effect. But if we have a moral obligation to create a better world (universe), and a universe with more, happier people is better than a universe with fewer, less happy people, then the total value realized by surviving long enough to escape the inner solar system and disperse throughout the galaxy is infinitely greater than a wholly Earthbound human existence, dwarfing any short-term goods that can only be achieved at the expense of this long-term future.

Homo sapiens is now well past the dangers of clade birth: humans and their animal domesticates now comprise a gigantic fraction of vertebrate biomass (~99%) and inhabit

nearly every ecological zone on Earth. The planet has been radically reconstructed to support human lifeways, a transformation that is evident from space. The immediate risk of human extinction is therefore small relative, say, to prehistoric hunter-gatherers or other medium-sized extant animals. Nevertheless, *Homo sapiens* is still in its technological infancy, and it is far from guaranteed that it will weather the volatility of its technological adolescence. This cautionary message is not new: it has been propounded for decades by space exploration theorists like Carl Sagan, Frank Drake, Stephen Hawking, and others who came of intellectual age during the brinkmanship of the Cold War when nuclear annihilation was a salient prospect. What is new about longtermism is that it represents a systematic attempt to integrate visions of a cosmic human future into an ethical framework for practical decision-making and policy-setting under conditions of moral risk and uncertainty.

Longtermist arguments repose on several controversial normative premises. First, they assume that a universe containing a greater number of morally valuable lives is better than a universe containing fewer (or no) morally valuable lives. Second, they assume that we have a moral duty to promote (maximize?) the goodness of future worlds, which, given the first premise, entails that we create larger numbers of value structures in the universe. This leads to the conclusion that we ought to prioritize policies that safeguard the long-term prospects of humanity over those that provide shorter-term goods that do not contribute to, or come at the expense of, our deep human futures. Third, many longtermists argue that postponing the extinction of human persons and their descendants is the best thing for the axiological future of life in the universe, even if doing so comes at the expense of sentient life more broadly on Earth as humans commandeer an ever larger fraction of energy in the biosphere in the service of making more human persons.

These are all contestable propositions. Perhaps welfare consequentialism is wrong (Heikkinen 2022) and we in fact have no strong moral obligation to promote aggregate value structures in the universe. As several authors have noted, the duty to ensure that any people we create are happy is very different, and far more plausible, than the duty to create happy people (see, e.g., Setiya 2014). And perhaps the value gains from creating persons (roughly, beings with sentience + rationality) do not outweigh the value losses occasioned by the extinction of much larger aggregates of sentient nonpersons they will inevitably bring about—and thus the living universe would be overall better off for our demise given the selfish, destructive, and violent apes that we are (or at least that half of *Homo sapiens* is!). Even if scientific rationality were the only conduit for promoting sentient life of any sort beyond its planet of origin, perhaps the longtermist focus would be better placed on disseminating loci of moral value, rather than persons per se.

Still, one would be hard-pressed to reject the idea that we ought to consider the effects of our current activities on future generations and, in addition, that these future-oriented considerations ought to have some moral teeth. If thorny 'Parfitian' problems in population ethics do not impede moral duties to future generations in the context of climate change, then it is not much of a stretch to think that they allow for duties to long-term future people. On the other hand, if longtermist assumptions should prove incorrect, then longtermist policies could come at a significant moral cost. For instance, longtermism seems to require that we prioritize the reduction of existential risks, even those with very small probabilities, over garden variety institutional reforms that will clearly better the lives of existing and soon-to-exist people (MacAskill 2022). Even more worrisomely, it appears to recommend

policies expected to secure small benefits to huge numbers of future people over policies expected to bring large benefits to the much smaller number of people living today, including the worst off among us. This raises the quite reasonable concern that merely *conjured* futures will justify inattention to current injustice and reign tyrannically over concrete lives of the present.

2 The epistemic challenge

2.1 Science fictiony futures

The biggest problem with longtermism, as I see it, relates not so much to the normative assumptions it makes, but to the predictive power it requires. I will assume for purposes of this chapter that strong longtermism is correct in broad moral strokes in order to train our sights on the key epistemic question: how 'long' can longtermism endeavor to be? The answer, of course, will hinge on the limits of human prognostication. If we knew with certainty that humans will be wiped off the face of the Earth within 200 years unless we do x, y, and z, then we would not hesitate to come up with an x-y-z Marshall Plan to avert human extinction (though the recent dark satire *Don't Look Up* suggests that even the most blatant existential threats would not be enough to overcome tribalism and self-interest). In reality, humans have a very poor track record of value forecasting, making it easy for special interest groups and motivated reasoners to exploit margins of error as reasons for threat skepticism, which in turn justifies inaction or deprioritization.

But the psychological foibles of human reasoning are only one side of the epistemic challenge to longtermist goals. The other and far more recalcitrant problem relates to the unpredictable nature of the phenomena themselves. Even if policies were crafted by rational Bayesians who accepted that not preventing harms to distant future persons is just as *pro tanto* wrong as harms to present and near-term persons (but see Lloyd 2021 for an opposing argument), it would still be rational for policy makers to discount harms to future people *epistemically*, as the visions painted by our keenest faculties of foresight grow blurry and fade to black over deep stretches of time. The very real possibility—and perhaps very high probability—that nobody will be around to benefit from the long-term future should surely count against longtermist policies in the competition for limited resources.

Following Tarsney (2022), I will refer to this as the 'epistemic challenge' to longtermism. The problem is even worse than a technical scenario of *uncertainty*: for not only are we ignorant of the probability distributions of morally relevant long-term futures that would allow us to manage moral risk, we do not even know what the relevant long-term futures might look like. Herein lies what Rini (2022) maintains is longtermism's 'fatal flaw':

> The problem is incredibly simple: we have no idea what the future will be like. Our trying to anticipate the needs of star-faring people millions of years hence is no more realistic than Pleistocene chieftains setting aside sharpened flint for twenty-first-century spearheads.

This criticism is not entirely fair, however. What longtermists are setting aside for preservation is not the modern equivalent of prehistoric spearhead materielles, but the continued existence of the species itself. This need not entail any claims about the specific conditions under which distant-future humans will flourish, apart from the condition that they come to exist.

Furthermore, although the cosmic visions of longtermism are fantastical, they are not fantasy. Legendary science fiction novelist Arthur C. Clarke remarked that any sufficiently advanced technology would be indistinguishable from magic. The key difference between science fiction and fantasy, however, is this: whereas fantasy depicts fantastical phenomena while black-boxing their causes under the supernatural catchall 'magic', science fiction explores the natural mechanisms underlying phenomena that seem fantastical from our blinkered perspective. Still, the fact that space colonization is scientifically plausible does not make it likely. If the aims of longtermism are to be genuinely long-term aims, then we must be able to make meaningful forecasts about the *distant* future.

'Distant' means different things to different people. For the average elementary-school-age child, the idea of having children may seem sufficiently remote to ignore. For many adults, distant futures are on the order of hundreds of years. Even decorated futurists are infamous for depicting social and technological worlds only decades or centuries out, replete with androids, daily-use jetpacks, flying cars, routine nuclear fusion, long-distance space travel, and the like. The philosophically contemplative universe of *Star Trek* puts humans at trans-galactic travel after extinguishing war, poverty, tribalism, and inequality all by the year 2265! Even Isaac Asimov, one of the greatest science fiction minds of the 20th century, predicted that humanity would be a full-fledged space-faring civilization by 2019; at the time of this writing, it is 2023 and humans have not set foot on the moon since Don McLean's folksy *American Pie* topped the U.S. billboard charts.

For the macroevolutionist who studies large-scale patterns of life on Earth, 'distant' connotes an entirely different order of temporal magnitude. It means contemplating living worlds separated from our own by tens of millions, hundreds of millions, and even thousands of millions of years. I will call these *macroevolutionary futures*—vast spans of time during which biological taxa originate, go extinct, and are succeeded by entirely different faunal assemblages. It is this geologically deep time frame that distinguishes truly long-term longtermism from standard approaches to intergenerational ethics, and it is against these deep temporal scales that the disvalue of human extinction is gauged.

The problem, however, is that once we insist on more than Pascalian probabilities for these macroevolutionary futures, the epistemic challenge becomes harder to overcome. Lacking the psychomathematics of Asimov's (1951) *Foundation*, we have no way of putting meaningful chances on galactic colonization or cosmic niche construction carried out by humans or their descendants (whether carbon-based, silicon-based, or in some other form). As magnificent as these conjured futures are to science fiction buffs (myself included), they border on the ludicrously anthropocentric, drastically overestimating humanity's agency in the face of the vast timescales and complexly configured forces that have shaped, and will continue to shape, the future of life on Earth. The apparent fact that no species in the history of the observable universe has achieved the kinds of cosmic feats that longtermists contemplate suggests that our own probability of achieving them is vanishingly low, perhaps to the point of being safely ignored for policy purposes (but see Ćirković 2018).

2.2 From natural history to unnatural futures

What, then, can we say about macroevolutionary futures and how they bear on the longtermist project? Anthropocentric assumptions about evolution, combined with the tendency to overlook observer-selection effects (Bostrom 2002), make it easy to slide from what happened here (on Earth) and to us (humans) to what has to happen everywhere and to everyone. This uncritical projection of Earthly evolution into the cosmos has led many astronomers, steeped in the invariance of physical law, to expect a galaxy awash in extraterrestrial civilizations, only to be greeted by cosmic crickets.

In order to weigh the expected disvalue of human extinction, we need to have a sense of what evolution has in store not only for our own species, but also for the animalosphere writ large. This, in turn, requires that we consult the deep past. As Sir David Attenborough remarks in his beautiful autobiographical film *A Life on Our Plant* (2020), natural history is not merely a historical science—it is also a study of the present and future of the biosphere. Some of the same investigatory tools that allow us to reconstruct the deep past also enable us to project into the future (Archer 2016; Currie 2018). For example, paleoclimatology uses proxy data in the geological record for parameters such as temperature, moisture, CO2 levels, and albedo to reconstruct ancient climates. From these reconstructions, we can gain insights into climate dynamics, fine-tune our climate simulations, and identify paleo analogs to our present climate predicament—all of which can inform predictions about future climate scenarios (Tierney et al. 2020; Lear et al. 2021; Watkins 2024). One key difference, however, between the past and the future is that humans cannot manipulate the past whereas they can, through their intentional (and unintentional) actions, change the future (Turner 2007). This 'unnaturalness' due to the input of human agency makes extrapolating from the deep past an especially risky epistemic enterprise.

3 Evolutionary progressivism

3.1 The *scala naturae* naturalized

If evolution were to drive teleologically toward certain predefined endpoints, and if we could infer those endpoints from long-term historical trajectories of life on Earth, then we would be well-positioned to make informed projections about macroevolutionary futures that could then be fed into longtermist arguments. The notion that the history of life drives toward functionally and morally better outcomes, culminating in human beings, is deeply embedded in the human psyche, codified as it was in the religiously inspired *scala naturae* that reigned in different guises for millennia. The *scala naturae* was thoroughly exploded on two separate fronts of the Darwinian assault on special creation. First, Darwin provided a mechanism—blind variation and natural selection—that could explain the origins of functional complexity without postulating any mysterious Lamarckian tendencies toward perfection. Second, Darwin made a persuasive case for the theory of common descent, an inference to the best explanation that framed humanity as simply one twig on an arborescent tree of life, rather than the anointed telos of a linear, progressive trend.

The notion that the history of life amounts to an inexorable march from 'monad to man' did not die, however, with the ascent of the Darwinian paradigm (Ruse 2009; O'Malley

and Powell 2016; Powell and Mikhalevich 2023). Progressivist assumptions continue to produce overconfident assertions about the likelihood of stellar conquest carried out if not by humans, then by their intellectually and morally superior successors such as postpersons or superintelligences, who sit closer than humans to angels in the naturalized *scala naturae*.

3.2 Complexification trends

Nevertheless, progressivist readings of evolution are not entirely without cause. From simple cells, evolution has produced complex cells; from free-living complex cells, evolution has produced aggregates of specialized cells that make up the highly differentiated bodies of complex multicellular organisms like animals, plants, and fungi; and from free-living complex multicellular organisms, evolution has produced highly specialized aggregates of complex multicellular organisms that comprise societies or colonies. This trend in increasing maximum 'nestedness' over the history of life has been vindicated scientifically (McShea 2001), and in theory, similar long-term trends could be documented for other dimensions of complexity, such as anatomical, metabolic, or cognitive complexity, were these phenomena amenable to scaling (a big 'were').

Yet even if there are long-term trends in complexity maxima, this does not imply there are built-in biases that would allow us to ascribe complexification as a 'goal' of macroevolution. Trends in complexity may simply be due to passive diffusion away from a minimum boundary (Gould 1997), with simplifications just as or more likely than step-ups in complexity throughout the history of life (O'Malley and Powell 2016). In the absence of systematic biases, there is no reason to suppose that evolution is driving human societies toward the assembly of a higher-level collectivity, or that life on Earth will ultimately consolidate in a single, planetary organism of the sort depicted in Stanislaw Lem's *Solaris* (for a discussion, see Ćirković 2018). Though titillating, such ideas germinate out of progressivist readings of evolution for which there is little theoretical support.

4 Life in a nomological vacuum

4.1 The generic nature of biological laws

In order to predict far-future states of any complex system—such as the position of celestial bodies in the solar system—that system must be governed by spatiotemporally invariant, epistemically accessible laws of nature. And the problem is that biology operates in what is largely a nomological vacuum. As Beatty (1995) influentially showed, nearly all distinctively biological regularities are accidents of an evolutionary process that could have gone in any number of possible directions due to fluctuating selection pressures, the random ordering of mutations, and drift events. These processes are quintessential prediction defeasors, rendering even the most successful biological generalizations spatiotemporally restricted and riddled with exceptions. For these reasons, laws of life will either be framed in such a way that they embed accidental antecedents (e.g., Sober 1997), or else lack the specificity needed to predict long-term evolutionary futures.

Take, for example, the Principle of Natural Selection, which is essentially a 'schema' that tells us what to expect when one trait has been determined to be fitter (i.e., to cause greater expected reproductive success) than another in a specified selective environment (Brandon 1991). There are no substantive laws of fitness that tell us *ex ante* which traits are the adaptive ones: relative fitness is determined by consulting a highly localized set of abiotic, biotic, and developmental factors (Rosenberg 1985). In some environments, an 'engineering analysis' will suggest that flexible intelligence courtesy of an elaborated nervous system is an adaptive trait; in others, rigid instinct operating with lower metabolic demands may be the superior survival strategy. As Lewontin (1978) points out, even in highly circumscribed situations, engineering analyses are nearly always incomplete. And even where we get the engineering analysis right, relative reproductive tendencies can always (literally, always—see Brandon 2006) be nullified by stochastic processes such as drift, mutation, environmental fluctuations, strategic biotic interactions, proverbial lightning strikes, and so on.

Consider another contender for a fundamental biological law: the so-called Zero-Force Evolutionary Law (ZFEL), proposed by McShea and Brandon (2010). ZFEL describes the default tendency of diversity and complexity to increase on average in the absence of selection, other forces, and constraints. Although ZFEL predicts that complexity will tend to increase over the long run, it is operating on a technical, non-intuitive notion of 'pure complexity' that does not map on to adaptation, functional integration, energy usage, computational power, cognitive flexibility, nestedness, or any other colloquial dimension of complexity. Insofar as longtermism is concerned with value structures generated by these latter types of complexity—such as sentience, personhood, and other embodied psychological processes arising out of specific body-brain configurations—the ZFEL, like other laws of life, does not provide the predictive precision that longtermism requires.

5 Natural history is not uniform

5.1 Extinction and origination

Charles Lyell, the founder of modern geology and a mentor of Darwin's, is credited with establishing the principle of *uniformitarianism*, which holds that the most fruitful way to understand events in the deep past is by extrapolating presently operating causes at their current rates and intensities. Yet even in Lyell's day, it was widely understood that geological processes were far from uniform (Gould 1995; Erwin 2011). As we shall see, this nonuniformitarian picture of terrestrial evolution, bolstered by decades of work in paleobiology, is critical to long-term value projections.

Perhaps the most glaring feature of the nonuniformitarian landscape is the phenomenon of mass extinction: great perturbations of the biosphere in which large fractions (70–95%) of the Earth's taxa vanish, often too quickly for their decline to be recorded in the strata. These patterns are not artifacts of a spotty fossil record, as Lyell and Darwin had surmised, but genuine ecological upheavals precipitated by a wide range of abiotic and biotic factors, from asteroid impacts and massive vulcanism to the bacteria-driven oxygenation of the atmosphere. What do patterns of mass extinction tell us about the sorts of futures that longtermists value?

First, they allow us to identify correlates of mass extinctions that can be used to assess the probability of a sixth mass extinction precipitated by human activities. These correlates include not only significant species loss, but also the mass-rarity of once abundant ecosystem engineers that decline below a critical ecological threshold to the point that they can no longer provide the ecosystem services that underpin biodiverse communities (Hull, Darroch, and Erwin 2015). For instance, the decline of foundational coastal coral reef and planktonic communities are historically linked to ecosystem collapse, and so the stunningly swift destruction of coral reefs all around the world due to warming, pollution, and overfishing should set off alarm bells (Pandolfi and Kiessling 2014). Yet here, too, major epistemic difficulties arise. The data used to compare present rates of species loss and decline with those of the big five mass extinctions are by necessity of an entirely different nature, making past and present rates of loss difficult to compare and calibrate (Bocchi et al. 2022; Bocchi 2022).

Second, the dynamics of mass extinction make predictions about long-term patterns of faunal turnover all but impossible and obliterate any possibility for evolutionary progress. Mass extinctions are partly random kill events that result in the wanton destruction of a large fraction of the Earth's biota, and partly selective sampling events that briefly impose different rules of survivorship than obtain during long stretches of halcyon time (Benton et al. 1996; Gould 2002). In either case, the taxa that emerge on the other side of hell are largely the children of fortune, not of comparative adaptive merit in any global sense. Furthermore, we have no way of predicting what will ensue once the biosphere recovers from a mass extinction. We cannot know *ex ante* which lineages will squeak through the boundary or what they will do with their eco-spatial inheritance when they do—and so we have no idea whether the next faunal regime will be morally better, worse, or similar (all things considered) than the one that preceded it. This is in part because, although mass extinctions are tremendously disruptive, they also play a role in evolutionary diversification by breaking up long-running structures of inter-clade suppression.

For instance, the end-Devonian crisis resulted in the severe depletion of placoderm (armored) fish, the dominant apex predators of the day, which were replaced by previously subordinate cartilaginous and bony fish. Likewise, throughout the Permian, archosaurs (a group that includes extinct dinosaurs and pterosaurs as well as living crocodiles and birds) were suppressed by a diverse group of mammalian ancestors called therapsids, until The Great Dying at the end of the Permian—the most destructive event in the history of animal life—killed off the therapsids, paving the way for crocodilian dominance in the Triassic. Dinosaurs, in turn, were the unlikely heirs of the Mesozoic, having themselves been suppressed by crocodyliforms until the end-Triassic mass extinction (Brusatte et al. 2008). And finally, mammals were confined to a largely nocturnal scavenging niche during the 170-million-year reign of the dinosaurs, until that reign ended abruptly on one very, very bad day that closed out the Mesozoic Era.

The unlikely radiation of the mammals—and by implication our own primate lineage—hinged on a 11km-wide bolt from the black that struck the Earth in just the right place, at just the right time, and at just the right angle to doom dinosaur ecosystems and tip the Earth's climate into a more mammal-friendly regime. The dinosaur and archosaur-dominated ecosystems of the Mesozoic were just as ecologically diverse, metabolically active, socially complex, and stunningly beautiful as our mammal-dominated era of evolution, if not more so. The recent BBC series *Prehistoric Planet* (2022), done in quintessential Attenborough style,

shows just how modern, familiar, and colorful—both metaphorically and literally—life was in the Cretaceous. This was not a primitive world destined for the dustbin of natural history, just as ours is not a superior world destined to last in perpetuity. Brutus was wrong: all of our days are numbered, but not for any fault of our own.

Another striking pattern in the fossil record is that species appear instantaneously and fully formed and do not change appreciably until they vanish from the rocks—a phenomenon that Eldredge and Gould (1972) dubbed 'punctuated equilibria'. As with mass extinction, the pattern of punctuated equilibria is not an artifact of imperfect preservation. Rather, it reflects the dominant mode of speciation recorded literally in the geological record. This nonuniformitarian pattern is not limited to speciation: as it turns out, the rapid origin of innovation followed by long-term stasis is borne out at the largest scales of life (Erwin 2011; Hughes, Gerber, and Wills 2013). This is perhaps most poignantly illustrated by Gould's 'decimation-diversification' hypothesis (1990; 2002): the idea that animal life began with a wide breadth of functional designs that was winnowed down over the course of the Phanerozoic Eon (~550mya–present). As deep time marched on, mass extinctions would repeatedly lop off large branches of the animal tree, leaving increasingly large and unbridgeable gaps between the remaining branches. The result is the inhomogeneous (clumpy) occupation of 'morphospace' that we see today, rather than an insensible gradation of forms. But perhaps, one might argue that instead of a gradual climb up the evolutionary ladder of progress, natural history is more like a ratchet of functional improvement with fits and starts between the cranks. Let us then look in more detail at the *causes* of these nonuniformitarian patterns.

5.2 Contingency and constraint

The contingent view of macroevolution can be encapsulated in two words: history matters (Beatty 2006; Powell 2012; Turner 2007; Powell and Mariscal 2015). The basic idea is that the direction life takes, in broad strokes and wherever it evolves, will depend on a highly unrepeatable series of events that could easily have been otherwise. In other words, there is no necessity to the shape of life as we know it, and so we must appeal to narrative rather than nomological explanation to describe these contingent counterfactuals (Beatty 2016; 2017). Through narratives we can draw upon a bounty of epistemic tools to piece together the quirky, Rube Goldberg-like chain of events that led to where we are today, including an analysis of plausible but unactualized historical paths. No matter how detailed these reconstructions, however, they will not yield any contentful predictions about the contingent confluence of causes that will conspire to construct our macroevolutionary future. Causes must precede their effects, and so while causal interactions in the present can change the shape of the future, they cannot influence events that have already occurred. We can therefore produce exquisite reconstructions of the deep past, while we are left with a clunky, course-grained ability to predict the macroevolutionary future.

Contingency alone does not complete the case against predictable progressivism, however. What makes contingent events like the patterns of early animal extinction so formative is that they become locked in place by developmental constraints (Powell 2020). In the initial phases of animal evolution, morphological design was still pliable and had the potential to be channeled in many different directions, as evinced by the menagerie of science

fictiony forms that inhabited the shallow seas of the early Cambrian. However, this pliability was short-lived: once the overarching developmental parameters of higher taxa were laid down, they became recalcitrant to re-design due to the topology of gene-regulatory networks.

In complex multicellular organisms, ontogeny begins with a fertilized egg and ends with the highly differentiated tissues and integrated organ systems that comprise the mature organism. Once 'upstream' nodes of these developmental cascades were connected up with 'downstream' batteries of genes that guide more fine-grained differentiation of organs, early-acting genes could no longer be modified without catastrophic consequences for the phenotype (Erwin 2011). The accidental survivors of early mass extinctions became *frozen* accidents, and the space of adaptive possibility was confined ever after to the parameters of surviving forms. Animal groups have diversified (and will continue to do so) within their own body plans: think of the astounding range of mammalian forms that arose with remarkable rapidity in the wake of the end-Cretaceous extinction, from bats and elephants to primates and whales. But the gaps between larger branches of the animal tree of life left by stochastic mass culling episodes are never 'regrown' due to the topographic constraints of development. If vertebrates went extinct, nothing functionally like them would ever arise again. This combination of contingency and constraint applies to trees of life everywhere and everywhen, as it is the source—not a product—of contingent processes on Earth.

5.3 Convergence and singularity

Some evolutionists have pushed back against the contingent view of life by arguing that the ubiquity of *convergent evolution*—the repeated origination of biological forms and functions—indicates that certain forces have operated consistently and efficaciously to drive trends that might be described as lawlike and even progressive (Conway Morris 2003; Beatty 2006; McGhee 2011). The convergentist argument against contingency goes roughly like this. Insofar as we are working with an 'N' of 1—a single history of life—the contingency question appears to be insoluble. But in fact, convergence is tantamount to natural experimental replication in the history of life. Impressive cases of convergence have been documented at all levels of life, from molecules and morphologies to the properties of mind. This ubiquity of evolutionary replication is evidence of a nomic undercurrent tugging animal evolution in particular and particularly progressivist directions, such as toward ever more sophisticated sensory modalities, information-processing capacities, social relations, and so on. The dynamics of contingency may temporarily set back these trajectories, but any such reversals of fortune will be short-lived (Vermeij 1999: 249) as evolution marches onward toward a more functionally complex and morally valuable world.

These conclusions simply outstrip the data. The causes of convergent evolution are heterogeneous, and many (probably most) cases of evolutionary repetition are due to shared developmental parameters that are themselves highly contingent (for an extended treatment of this issue, see Powell 2020). Basically, conserved internal parameters can taint the independence of natural experimental setups by rendering certain adaptive solutions more accessible to selection than others. Even where we have established genuinely robust cases of independent replication, we also need to consider their temporal distribution, as some

convergent outcomes may fail to re-emerge when they are extinguished in more advanced stages of life's sequence.

Consider, for instance, the cluster of coevolving traits that I have elsewhere called the 'welfare platform' (Powell 2020; Powell, Mikhalevich, and Buchanan 2021)—a set of interlocking sensory, neurological, cognitive, conative, and locomotory systems that plausibly give rise to beings with moral standing. There is strong evidence that the welfare platform arose independently in arthropods, vertebrates, and mollusks (in that order) all within a relatively short span of time during the early phases of animal evolution. As far as we can tell, the welfare platform never arose again in the 0.5 billion-year history of animal life on Earth, perhaps because to do so would require the radical reorganization of body plans that the topography of developmental constraint does not permit (subsection 5.2). If the 'welfare platform' were to vanish at this late point, it is a fair bet that it would be gone for good—and with it, intrinsic value in the biosphere. Mental properties are the ethical crux of longtermist futures, and so it behooves us to delve more deeply into contingencies, repetitions, and irreversibilities in the evolution of minds.

6 The axiological irreparability of human extinction

6.1 Life's little joke

In his book *Full House*, Stephen Jay Gould (1997) notes an interesting irony: we often view the last surviving species of a once bushy clade as the pinnacle of ladder-like progress in their evolutionary line, rather than as the lingering vestiges of a once thriving animal group now teetering on extinction. Gould called this 'life's little joke'. Examples include rhinos and elephants—lineages whose extinction would take the entirety of a higher taxon with them. And the joke is on us, too. *Homo sapiens* is the last remaining twig of a once bushy, blooming, and buzzing branch of the primate order. Dozens of bipedal hominin species used to inhabit Africa and Eurasia, including varieties of australopithecines, habilines, erectus-grade humans, and numerous other *Homo* lineages; as of ~40,000 years ago, there is only one. Unlike mammalian megafauna, humans are not rare animals (at least not since the last super volcano eruption in Sumatra ~75,000 years ago), and it is far from clear that evolution is shuffling *Homo sapiens* to the exit (though we may promptly see ourselves out!). But the topology of life's little joke, combined with contingency and constraints in evolution, speaks to the irreparable moral catastrophe of human extinction. This leads us to a different irony, as the case against evolutionary anthropocentricism also implies that the extinction of human beings would be a moral travesty for the cosmos!

6.2 Humans vs. humanity

It is important here not to conflate the extinction of humanity as a biological or fossil species with the extinction of humanity understood in the functional Kantian sense as 'rational agency'. *Homo sapiens* could go extinct but not before giving rise to a daughter species comprised of equally morally valuable persons with equivalent (or greater) axiological

potential. It is the loss of certain mental properties—not the contingent taxa which instantiate them—that matters for longtermist arguments.

As noted earlier, although sentience looks to be a robustly replicable evolutionary outcome (judging by its independent realization in disparate animal phyla), it is unlikely to re-emerge at this mature juncture in the evolution of animal developmental systems. Thankfully, because sentience is found in extremely diverse animal groups distributed in all habitats on Earth, it is unlikely to be extinguished in the next mass extinction (anthropogenic or otherwise). We can rest assured that the welfare platform will persist, retaining a basis from which more sophisticated cognitive capacities could evolve, such as those that configure higher moral status.

Personhood (subjectivity + rationality) enjoys a far narrower phylogenetic distribution than sentience, though it has been realized several times in distant tetrapod vertebrates, including apes, whales, elephants, crows, parrots, and other plausibly self-aware species with sophisticated reasoning capabilities, stable psychological identities, emotional capacities, complex social structures, and so on. The repeated evolution of personhood within mammals, and more significantly, between mammals and birds (theropod dinosaurs)—lineages which diverged >300 million years ago and that have independently derived neuro-cortical structures—suggests that the loss of personhood is unlikely to be irreversible. There might be millions of years in lost value until the re-re-emergence of functional persons, but this would be a negligible moral hiatus in the grand scheme of life of Earth.

6.3 Humanity vs. technology

However, personhood in itself lacks the axiological potential on which longtermist arguments are predicated. It is only when personhood is combined with robust technological capacities that a high-value species can achieve a global distribution, comprise an ever-greater fraction of its planet's biomass, utilize an increasingly large proportion of energy flow in the biosphere, and eventually attain a meaningful presence in the cosmos (absent defeasors). Robust technological capacities, in turn, hinge on the evolution of *cumulative culture*, which allows for the scaffolding and step-wise improvement of knowledge, skills, innovations, and learning environments down the generations (Tennie, Call, and Tomasello 2009).

Cumulative culture is a shockingly recent evolutionary invention. It was absent for ~4 billion years on Earth until the arrival of *Homo*. It was ostensibly lacking in *Homo erectus* who, despite being a seafaring apex predator, robotically reproduced the same Acheulian hand axe without improvement for >500,000 years. Most puzzling of all, cumulative culture appears to have been lacking in anatomically modern *Homo sapiens* all the way up until the Upper Paleolithic, when human technology explodes onto the scene in its familiar modern forms. In other words, humans were persons for millions of years before they were capable of cumulative technological improvement. And no other animal persons have ever developed this capacity. This leaves a yawning gap between the evolution of personhood and the emergence of cumulative technology.

Within this gap lies the fortuitous confluence of various sophisticated cognitive abilities that underwrite cumulative culture. These include theory of mind, causal reasoning, mental

time travel, means-end rationality, metacognition, pedagogy, moral sentiments, language, imitation, joint intentions, and other faculties that underpin high-bandwidth/high-fidelity/high-innovation cultural transmission (Sterelny 2012). Each of these traits originated hundreds of thousands of years before the advent of cumulative culture, resting on layer upon layer of contingency and indicating a substantial (if currently unknown) threshold for cultural cumulativity. Many of these cognitive capacities arose in humans for their cooperative functions, such as the role they play in collaborative foraging, brooding, and information sharing. Yet the same cooperative feats can be achieved without any fancy cognition at all, as delightfully demonstrated by the social insects (Powell 2023). Further, this panoply of derived cognitive abilities in humans piggybacked on preexisting anatomical capacities, such as fine-motor manipulation, and they required an environment in which tools could be kinetically effective and thermal energy controlled, making the oceans favorable to the evolution of personhood (in whales) but unfavorable to cumulative culture. Finally, there was a substantial gap between the evolution of cumulative culture and the development of a scientific epistemology necessary (and sufficient?) to achieve longtermist goals. This ~60,000-year delay is likewise replete with contingencies, and it is why, as Jebari (2021) points out, human extinction may not be necessary for a morally catastrophic longtermist outcome. Underkill events that result in severe cultural setbacks could permanently impede technological maturity and thus be just as axiologically devastating as full-blown extinction. As with evolution in general (section 3), progress in cultural evolution is far from assured. There is no indication that adjacent *Homo* species would eventually have hit upon cumulative (let alone scientific) culture if humans had not beaten them to the punch. And if *sapiens*—the lone surviving member of their genus—were to go extinct, there would be no hominid platform for evolution to start from, making it unlikely that cumulative culture would ever arise again on Earth.

7 The epistemic challenge revisited

Let us now bring these paleobiological threads together and ask: does the contingent nature of natural history help or hurt the case for longtermism? The answer is that it does both. On one hand, the lack of specific laws of fitness, the inherent unpredictability of macroevolutionary patterns and processes, and the plausible limitations of human agency in the face of geological forces make predicting human biocultural futures a dicey enterprise, to say the least. The chances that many millions of years from now the human lineage or its descendants will still be inhabiting the Earth or will have migrated to other worlds are entirely unknown and probably unknowable. It difficult, therefore, to know how much to discount deep-future people in our ethical decision-making.

On the other hand, a lot more rides on human extinction in an evolutionary world sprinkled with singularities than it does in one riddled with repetitions. If there is no law-like necessity to the twists and turns that led to the emergence of humanity in the functional sense, then the loss of *Homo sapiens* and its descendants is likely to be evolutionarily irredeemable. When we fold in the absence of detectable galactic-level civilizations over the last 13 billion years of cosmic evolution, human extinction begins to look like an axiological catastrophe of astronomical proportions, for it suggests that no other species is likely to fill the terrestrial, let alone galactic, moral void anytime soon.

Truly long-term longtermism will remain beyond our moral horizon for the time being, and perhaps incurably so. But thinking bigger is not always better. Medium-term longtermism, which would require (e.g.) that we project tens or even hundreds of millennia into the future, is probably within our horizon of epistemic plausibility. Fortunately, many of the mitigation strategies that reduce existential risk—such as slowing climate change, easing poverty, and managing intergroup conflict—benefit presently existing, near-term future, and medium-term future people alike. To this extent, longtermism recommends that we do what we should have been doing anyway, and so one struggles to see what actionable items it adds to our already packed moral agenda (Rini 2022). Personally, I am at peace with the idea of a terrestrial and cosmic future without human beings, and I am not convinced that populating the universe with quadrillions of jovial humans is morally required or even desirable. And yet, I find myself agreeing with the longtermist sentiment—perhaps for sentimental reasons—that we should do everything in our power to stave off the conclusion of the human story, even if sooner or later, for better or for worse, it will come to an end.

Acknowledgments

I am grateful to Aja Watkins, Federica Bocchi, and Irina Mikhalevich for their comments on an earlier draft of this manuscript, and to the folks at the Global Priorities Institute for bringing me into this fascinating and important discussion.

References

Archer, D. (2016), *The Long Thaw: How Humans Are Changing the Next 100,000 Years of Earth's Climate* (Princeton University Press).

Asimov, I. (1951), *Foundation* (Gnome Press).

Beatty, J. (1995), 'The Evolutionary Contingency Thesis', in G. Wolters and J. Lennox (eds.), *Concepts, Theories, and Rationality in the Biological Sciences* (University of Pittsburgh Press), 45–82.

Beatty, J. (2006), 'Replaying Life's Tape', in *Journal of Philosophy* 103: 336–362.

Beatty, J. (2016), 'What Are Narratives Good For?', in *Studies in History and Philosophy of Biological and Biomedical Sciences* 58: 33–40.

Beatty, J. (2017), 'Narrative Possibility and Narrative Explanation', in *Studies in History and Philosophy of Science Part A* 62: 31–41.

Benton, M. J., Jablonski, D., Erwin, D. H., and Lipps, J. H. (1996), 'On the Nonprevalence of Competitive Replacement in the Evolution of Tetrapods', in J. W. Valentine and D. Jablonski (eds.), *Evolutionary Paleobiology* (University of Chicago Press), 185–210.

Bocchi, F. (2022), 'Biodiversity vs Paleodiversity Measurements: The Incommensurability Problem', in *European Journal for Philosophy of Science* 12/64.

Bocchi, F., Bokulich, A., Castillo Brache, L., Grand-Pierre, G., and Watkins, A. (2022), 'Are We in a Sixth Mass Extinction? The Challenges of Answering and Value of Asking', in *British Journal for Philosophy of Science*. https://doi.org/10.1086/722107

Bostrom, N. (2002), *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge).

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

Brandon, R. N. (1991), *Adaptation and Environment* (Princeton University Press).

Brandon, R. N. (2006), 'The Principle of Drift: Biology's First Law', in *The Journal of Philosophy* 103/7: 319–335.

Brusatte, S. L., Benton, M. J., Ruta, M., and Lloyd, G. T. (2008), 'Superiority, Competition, and Opportunism in the Evolutionary Radiation of Dinosaurs', in *Science* 321/5895: 1485–1488.

Buchanan, A. (2020), *Our Moral Fate: Evolution and the Escape from Tribalism* (MIT Press).

Buchanan, A. and Powell, R. (2016), 'A Naturalistic Theory of Moral Progress', in *Ethics* 126: 983–1014.
Buchanan, A. and Powell, R. (2018), *The Evolution of Moral Progress* (Oxford University Press).
Ćirković, M. M. (2018), *The Great Silence: Science and Philosophy of Fermi's Paradox* (Oxford University Press).
Conway Morris, S. (2003), *Life's Solution: Inevitable Humans in a Lonely Universe* (Cambridge University Press).
Currie, A. (2018), *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences* (MIT Press).
Eldredge, N. and Gould, S. J. (1972), 'Punctuated Equilibria: An Alternative to Phyletic Gradualism', in T. J. M Schopf (ed.), *Models in Paleobiology* (Freeman Cooper), 82–115.
Erwin, D. H. (2011), 'Evolutionary Uniformitarianism', in *Developmental Biology* 357/1: 27–34.
Gould, S. J. (1990), *Wonderful Life: The Burgess Shale and the Nature of History* (W. W. Norton & Company).
Gould, S. J. (1995), 'Tempo and Mode in the Macroevolutionary Reconstruction of Darwinism', in *Proceedings of the National Academy of Sciences* 91/15: 6764–6771.
Gould, S. J. (1997), *Full House* (Harvard University Press).
Gould, S. J. (2002), *The Structure of Evolutionary Theory* (Belknap Press).
Greaves, H. and MacAskill, W. (2021), 'The Case for Strong Longtermism', GPI Working Paper No. 5-2021 (Global Priorities Institute, Oxford University).
Heikkinen, K. (2022), 'Strong Longtermism and the Challenge from Anti-aggregative Moral Views', GPI Working Paper No. 5-2022 (Global Priorities Institute, Oxford University).
Hughes, M., Gerber, S., and Wills, M. A. (2013), 'Clades Reach Highest Morphological Disparity Early in Their Evolution', in *Proceedings of the National Academy of Sciences* 110/34: 13875–13879.
Hull, P. M., Darroch, S. A., and Erwin, D. H. (2015), 'Rarity in Mass Extinctions and the Future of Ecosystems', *Nature* 528/7582: 345–351.
Jebari, K. (2021), 'Replaying History's Tape: Convergent Cultural Evolution and the Prospects of Humanity after a Social Collapse' (SSRN 3840244).
Kumar, V. and Campbell, R. (2022), *A Better Ape: The Evolution of the Moral Mind and How it Made Us Human* (Oxford University Press).
Lear, C. H. et al. (2021), 'Geological Society of London Scientific Statement: What the Geological Record Tells Us about Our Present and Future Climate', in *Journal of the Geological Society* 178: jgs2020-239.
Lewontin, R. C. (1978), 'Adaptation', in *Scientific American* 239/3: 212.
Lloyd, H. R. (2021), 'Time Discounting, Consistency, and Special Obligations: A Defence of Robust Temporalism', GPI Working Paper No. 11-2021 (Global Priorities Institute, Oxford University).
MacAskill, W. (2022), *What We Owe the Future* (Basic Books).
McGhee, G. (2011), *Convergent Evolution: Limited Forms Most Beautiful* (MIT Press).
McShea, D. W. (2001), 'The Hierarchical Structure of Organisms: A Scale and Documentation of a Trend in the Maximum', in *Paleobiology* 27/2: 405–423.
McShea, D. W. and Brandon, R. N. (2010), *Biology's First Law: The Tendency for Diversity and Complexity to Increase in Evolutionary Systems* (University of Chicago Press).
O'Malley, M. and Powell, R. (2016), 'Major Problems in Evolutionary Transitions: Toward a Metabolism-Based Account of Macroevolution', in *Biology and Philosophy* 31/2: 159–189.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Pandolfi, J. M. and Kiessling, W. (2014), 'Gaining Insights from Past Reefs to Inform Understanding of Coral Reef Response to Global Climate Change', in *Current Opinion in Environmental Sustainability* 7: 52–58.
Persson, I. and Savulescu, J. (2012), *Unfit for the Future: The Need for Moral Enhancement* (Oxford University Press).
Powell, R. (2012), 'The Future of Human Evolution', in *British Journal for the Philosophy of Science* 63: 145–175.
Powell, R. (2020), *Contingency and Convergence: Toward a Cosmic Biology of Body and Mind* (MIT Press).
Powell, R. (2023), 'Social Norms and Superorganisms', in *Biology &* Philosophy 38/3: 1–25.
Powell, R. and Mariscal, C. (2015), 'Convergent Evolution as Natural Experiment: The Tape of Life Reconsidered', in *Interface Focus* 5/6: 20150040.
Powell, R. and Mikhalevich, I. (2023), 'Wonderful Mind', *Journal of the Philosophy of History* 17/1: 77–103.
Powell, R., Mikhalevich, I., and Buchanan, A. (2021), 'How the Moral Community Evolves', in J. Savulescu and S. Clarke (eds.), *Rethinking Moral Status* (Oxford University Press).
Rini, R. (2022), 'An Effective Altruist? A Philosopher's Guide to the Long-Term Threats to Humanity', in *Times Literary Supplement*.
Rosenberg, A. (1985), *The Structure of Biological Science* (Cambridge University Press).
Ruse, M. (2009), *Monad to Man: The Concept of Progress in Evolutionary Biology* (Harvard University Press).

Setiya, K. (2014), 'The Ethics of Existence', in *Philosophical Perspectives* 28/1: 291–301.
Sober, E. (1997), 'Two Outbreaks of Lawlessness in Recent Philosophy of Biology', in *Philosophy of Science* 64/S4: S458–S467.
Sterelny, K. (2012), *The Evolved Apprentice* (MIT Press).
Tarsney, C. (2022), 'The Epistemic Challenge to Longtermism', GPI Working Paper No. 3-2022 (Global Priorities Institute, Oxford University).
Tennie, C., Call, J., and Tomasello, M. (2009). 'Ratcheting Up the Ratchet: On the Evolution of Cumulative Culture'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364/1528: 2405–2415.
Thorstad, D. (2021), 'The Scope of Longtermism', GPI Working Paper No. 6-2021 (Global Priorities Institute, Oxford University).
Tierney, J. E. et al. (2020), 'Past Climates Inform Our Future', in *Science* 370/6517.
Turner, D. (2007), *Making Prehistory: Historical Science and the Scientific Realism Debate* (Cambridge University Press).
Vermeij, G. J. (1999), 'Inequality and the Directionality of History', in *The American Naturalist* 153/3: 243–253.
Watkins, A. (2024). 'Using paleoclimate analogues to inform climate projections'. in *Perspectives on Science*, 32(4), 415–459.

12
Coping with Myopia

Philip Kitcher

> . . . the rope consists of fibres, but it does not get its strength from any fibre that runs through it from one end to the other, but from the fact that there is a vast number of fibres overlapping.
>
> Wittgenstein, *The Blue and the Brown Books*

During the past four decades, a growing number of people have come to recognize that our everyday actions are causing changes to our planet that are likely to pose significant challenges for our descendants. Indeed, the severity of the problem should, by now, be blindingly obvious to all. A sobering fact: more than half of the greenhouse gas emissions since the mid-eighteenth century (the pre-dawn of the industrial revolution) have occurred during the last 30 years (since 1990, the year of the first Intergovernmental Panel on Climate Change (IPCC) report[1]).

The problem of climate change is only one of several challenges that *potentially* call for long-term planning. Others include the dangers posed by the existence of nuclear weapons, pandemics, and the further development of artificial intelligence. Whether any of these will prove problematic in the long term is something we cannot yet know. In each instance, the future may lead us through a phase in which our descendants need to cope with the dangers, followed by a period in which the threat has been nullified. For the moment, however, we have no assurance of that. Hence, it seems wise to consider each problem as one that will persist indefinitely.

In what follows, I shall focus solely on the climate change problem. I do so for three reasons. First, I hold the metaphilosophical view that discussions of any of the long-term challenges ought to be informed by some serious knowledge of how the threat arises, and climate change is the example about which I know most. Second, at least two of the other problems—the challenges generated by pandemics and by nuclear weapons—are likely to be exacerbated unless the climate change problem is addressed; in a scenario Evelyn Fox Keller and I have outlined (Kitcher and Keller 2017: xii–xvi), future droughts make regions occupied by large numbers of people uninhabitable, leading to mass migrations that allow for viral mutations to invade human populations and that also provoke tensions with local inhabitants in the regions to which the migrants move; those tensions erupt into warfare, which escalates into an exchange of nuclear weapons. Third, and most importantly,

[1] https://ieep.eu/news/more-than-half-of-all-co2-emissions-since-1751-emitted-in-the-last-30-years, accessed 6 January 2025.

Philip Kitcher, *Coping with Myopia* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Philip Kitcher 2025. DOI: 10.1093/9780191979972.003.0012

I believe that the conclusions I shall draw about climate change offer a blueprint for addressing the other long-term challenges.

The scenario Keller and I have offered stops just short of human extinction. It thereby raises the possibility that a slightly intensified version might lead to a world in which our species has (prematurely) gone extinct. There are thus two related questions those who worry about the effects of our actions on the future of our planet must consider: What should we do to enable the human beings who come after us to have the opportunity to live acceptable lives? What should we do to enable our species to endure through an appropriate period? (As I have stated them, these questions urgently need specification. How should we understand 'acceptable' and 'appropriate'? This specification must, I claim, emerge from the deliberations of those who engage with the questions in their vague form.) In principle, it is possible for an answer to one of the questions to constrain the answer to the other. Imagine that planning aimed at enhancing the well-being of our descendants increases the chances of premature human extinction. For the purposes of this essay, however, I shall adopt a simplifying assumption: fulfilling the obligation to provide a habitable planet for our descendants does not raise the probability that human life will cease earlier than it otherwise would. With that assumption, we can explore the question of what we should do to provide for the people who will come after us, without worrying about the extinction issue.

Before taking up that question in earnest, one further clarification is needed. Is it correct to restrict attention to the *human* future? Many people (of whom I am one) believe that sentient non-human animals deserve our moral consideration. How do our obligations to those animals affect long-term planning? Do we have further obligations to preserve non-sentient aspects of the planet, to leave behind the forests and the mountains, as well as an ('appropriate') distribution of species to inhabit them?

With respect to the vast majority of inhabitants of the sentient province of the animal kingdom, we have only negative duties. In interacting with those animals, we should avoid harming them. There are a few animals whose lives we know how to improve, without inflicting any suffering on others. In the main, however, attempts at positive action inevitably play favorites. The idea of a Peaceable Kingdom, in which our efforts lead to the happiness of all, is a pre-Darwinian myth. Nature is not entirely red in tooth and claw, but competition is fierce enough to block attempts to increase universal well-being: give aid to predators and prey will suffer, and conversely. Natural selection makes Darwin's world almost completely ungovernable (for more detail, see Kitcher 2018).

On what moral grounds would we be required to leave the planet in some particular shape when human life is finally snuffed out? To preserve an environment in which animals we like (polar bears or nightingales, say) can thrive is to register a preference that no longer has any significance when *Homo sapiens* is no more. As far as I can tell, the only moral consideration for favoring particular characteristics for the post-human Earth lies in the possibility that some species of intelligent life might someday evolve, capable of benefiting from whatever of our achievements can be preserved and stored for the possibility that members of that species might have access to it. Unfortunately, what Francis Crick dubbed 'Orgel's second law' applies to that thought: we should remember that evolution is much cleverer than we are. We are completely clueless about what might foster intelligent life after our own demise.

Much more could be said about several of the simplifications I have made. They have enabled a focused question: Given the facts of global heating, what should we do to leave a habitable planet for our descendants, both in the immediate future and in the longer term?

Many of those who are aware of the causes and effects of climate change believe the question calls us to change our ways. Immediately and radically. They accept an obligation to abandon our practices of burning fossil fuels and pursuing industrial agriculture, in order to make the lives of the people who will come after us less harsh and precarious than they would otherwise have been. As I have noted, extending this obligation to encompass the welfare of non-human animals leads back to human well-being. The usual suspects are kinds of animals human beings happen to like—and can expect future human beings to want to have around. If there is a more abstract concern with species diversity, that too rests ultimately on the interests of future people, on the aesthetic joys and practical gains they may derive from a less spoilt and more varied world.

My aim, then, is to understand how we might determine the nature and scope of our responsibility for the environments future human beings will inhabit. The next step must address two major questions. First, who are these people for whose well-being we are responsible, and how far is the temporal distance between them and us? Second, how should we decide how to balance the sacrifices that must be made? To what extent ought the burdens fall on living people (on *which* living people?), and how should they be distributed across the centuries during which (with luck) human beings (and our evolved descendants?) will continue to live on our planet?

* * *

My discussion is guided by some epistemic assumptions and by an account of moral methodology I have offered elsewhere (in most detail in Kitcher 2021a; in application to the problems posed by climate change, in Kitcher and Keller 2017). Neither the assumptions nor the account will receive much defense here. A brief summary: I assume that our collective attempts at forecast are myopic—most (but not all) of our reliable predictions are focused on the short-term, on the condition of the planet roughly half a century from now; I also take moral progress to be promoted, not by consulting allegedly privileged texts or from individual efforts of ingenious philosophers, but by collective processes of deliberation guided by three ideals—the deliberations should be inclusive with respect to the perspectives of all who will be affected by the issue, they should be well-informed, and the participants ought to be committed to mutual engagement and to finding a solution all can live with.

If the issue is posed in terms of the responsibilities of the current human population to the same population in an undifferentiated future, we must start by considering which groups of people would be affected, for good or ill, by the actions available to us. Prominent among them, of course, are the younger members of the existing human population and successive generations of ever more remote descendants. Immediate and decisive curbing of emissions would benefit them by subjecting them to a lower global mean temperature. It might also bring disadvantages if the world economy collapsed, or if some resources we enjoy were permanently lost. The strong resistance to climate activism sometimes stems from viewing the changes demanded as inevitably bringing losses of these kinds, but, as

prominent economists have cogently argued, measures can be taken to avoid the dismal consequences.[2] The options for climate mitigation should thus be considered as packages of efforts at emissions reduction wrapped in socio-economic protections.

If no groups of people were adversely affected by pursuing options of this kind, decision-making about climate change would be easy. The familiar resistance to those options arises because there are groups who reasonably fear the impact of the emission-reduction strategy. In affluent nations, a sizeable segment of the population is wealthy enough to adapt to activist measures without significant hardship: people can pay higher costs for electricity generated without reliance on fossil fuels, can switch to a diet freed from the agricultural practices that contribute one third of global emissions, can replace gas-guzzling cars with electric vehicles (or use public transport), and so forth. Even in the rich nations, however, large numbers of people are unable to modify their lives in these ways. They cannot try to 'go green', unless they abandon their jobs and fall into extreme poverty, and they worry that destitution will be forced upon them. So a first constituency of negatively affected stakeholders consists of the *precarious poor*.

Around the world, many nations currently yearn for electrification. They need an infrastructure of the kind affluent countries take for granted, one that will supply and distribute energy wherever there is demand for it. In many instances, the cheapest and fastest route to their goal would be the one the world's original industrial giants (the UK, the US, and Germany) once adopted, and which later leaders in the global economy (China) have since imitated, namely to rely on native sources of fossil fuels. Without significant assistance from those responsible for the dangerous atmospheric concentration of greenhouse gases, most would-be developers cannot afford any alternative. A second group of stakeholders whose prospects would be dashed by rigorous climate policy are the *delayed developers*.

Both of these constituencies could be accommodated if the world's affluent nations were prepared to spend enough in compensation. Heavy investment in social programs to protect the precarious poor, together with massive infusions of aid to the delayed developers, would allow a pan-human alliance to make a quick transition to a world-wide green economy. Where are the funds to come from? Concerned about their future positions in a global capitalism involving large numbers of new competitors, the potential suppliers (wealthy countries) are reluctant to deplete their own economies. If they are not to fall behind, money must be diverted from 'unproductive sectors'—education must be focused on training the young to compete, expensive social and cultural 'luxuries' must be sacrificed. Pursuing that option would have negative effects on what matters most to the *cultural conservators*, people who see a sustainable future bought by abandoning social institutions and historical accomplishments as ransacking a treasury containing what gives irreplaceable values to human lives.

Thus arises a four-sided dilemma (a quadrilemma?). Our patterns of behavior have important consequences for all four of these constituencies: members of future generations, the vulnerable citizens of affluent nations, people in countries that yearn for development, and those who wish to preserve the social and cultural achievements of the past. Were the most powerful nations of the world to impose an immediate policy of drastically reducing greenhouse gas emissions, we might alleviate the plight of future people and preserve (at

[2] See Kitcher and Keller 2017, 115–23 and the references in the notes, 240–41.

least in the affluent world) the valuable heritage of the human past—but at the cost of further depressing the lives of the second and third groups. If we were to abandon many of the advances made possible by the industrial revolution (a Thoreauvian return to the simple life), we might sacrifice much that our predecessors have provided and 'level down' to a sustainable, egalitarian world. Or, we might opt for the policy on which the world seems currently to have decided—business as usual—and write off the human future.[3]

Even in this unnuanced version of the problem, in which future people are lumped together as a single class, there are four sets of stakeholders. Many philosophers will seek a principle apportioning burdens among these groups. They think of morality as somehow determinate, a matter for discovery, rather than, as I do, an evolving project. On my account, the moral project began, probably between 50,000 and 100,000 years ago, in response to a deep tension between our evolved psychological capacities and the form of society in which we live (Kitcher 2011). Like our biological cousins, the chimps and the bonobos, we have an adapted ability to identify the wishes and intentions of our fellows and to modify our behavior to promote their success—call this capacity 'responsiveness'. Responsiveness made it possible for us to live in relatively large societies (40–70 individuals) mixed by age and sex, but because of the limitations of our responsiveness to others, our social relations were constantly threatened. Our ancestors could live the kind of lives they did, but not easily. Fortunately, the human past has introduced the moral project, which has stepped in to amplify our adapted responsiveness. But situations constantly arise in which greater amplification is required. When that happens, the project needs extension. We have to develop morality further.

How? If the problem lies in the limits of evolved responsiveness, there's an obvious solution. Assemble representatives of the groups affected by the issue at hand. Inform them as thoroughly as possible about the factual details of the situation. Then encourage them to understand one another's predicaments and attitudes, to feel the world through others' skins, and to seek the closest approximation to accommodating the needs of all. The collective discussion I envisage doesn't produce outcomes that mirror some antecedent moral reality. The objectivity of any solution consists in its fit to the problem posed. We need keys to open crucial locks.

Given this approach, what's required for moral progress with respect to climate action is a deliberation among representatives of the four constituencies, one that's inclusive, well-informed, and mutually engaged. Yet, for two[4] obvious reasons, that might appear utterly impossible. The future generations are not here to participate. And our ability to fathom the future is obviously limited. As we'll now see, these two questions are interconnected.

* * *

My embryonic moral methodology has to cope with a number of instances in which potentially affected parties cannot speak for themselves. There are the very young, the developmentally disabled, the demented, and the non-human animals. The obvious solution is to involve people who understand the situations and interests of any absent class. Old people who can no longer think clearly are represented by those who know them best, and can

[3] The dilemma/quadrilemma whose provenance I have sketched here is discussed in far more detail in Kitcher and Keller (2017). See, in particular, 165–73, which draw on arguments developed in earlier discussions.

[4] A third, perhaps even more evident, charges that we could not realistically expect any such conversation to be fruitful. That concern will be considered later.

testify to what they intended for an end to their lives. Experimental animals can be represented by advocates who understand their normal mode of life, and their sensitivities to pleasure and to pain (Kitcher 2015). Extrapolating, we might call on representatives of future generations.

But who are these people yet unborn? When do they live, and what is it to represent them? Posing these questions leads into a familiar thicket: Derek Parfit's famous 'non-identity problem' (Parfit 1984), an issue on which a vast amount of philosophical ink continues to be spilt. The non-identity problem arises from the fact that our present actions not only (collectively) determine the circumstances in which our descendants will live, but also which individuals will exist. If we continue business as usual, one collection of human beings will live in the world of 2150. If we take large steps to curb emissions, the inhabitants of that world will be different (indeed, the intersection of the two sets will be empty). Hence, we can't be said to *harm* the people whose lives will be battered by the effects of business as usual. For, if we had amended our ways, they wouldn't exist.

Does this absolve us of all responsibility for future generations? Parfit didn't think so, and (to the best of my knowledge) neither do the many philosophers who have wrestled with his problem. The prevalent view takes the argument to offer a challenge: show how we can be under obligations to others to refrain from actions, even when those actions don't harm them.

I suggest a different way of conceiving the kind of obligation with which we are concerned in thinking about climate change. Start with a closely analogical case. A steward is appointed to maintain a public garden, set up to give pleasure to the people in a small town. The person who undertakes this role is lazy, and, after a short while, the flowerbeds are full of weeds, the grass has grown haphazardly, some trees have suffered from insects or from storms. Fewer people come, and those who do find the unkempt garden less inviting than they used, even a depressing place in which to linger. We *might* describe the steward's moral dereliction in terms of the harm visited on the visitors. But a better characterization is available. The steward hasn't done the job. The obligations of the role have been slighted.

Role obligations are typically directed towards people (or sometimes other animals) who satisfy a particular description. In the case of the steward, it's 'visitors to the garden' and it's not marked out in advance just who those visitors are. Similarly, veterinarians have obligations towards the pet-owners and the pets for which they seek treatment; mothers have obligations towards their children.

We *might* also describe maternal obligations using the vocabulary of 'harm'. The mother has a duty to nurture her child, a duty grounded in the fact that nonfulfillment would cause the child harm. We might also ask whether the obligation to nurture extends to a period of time before the child is born. Doesn't the pregnant woman have a duty to refrain from activities that might interfere with fetal development? Or, focusing on an even earlier time, is the duty in force even before the child is conceived?

Imagine a fun-loving young woman. She's the life and soul of many parties, often drinks to excess, smokes, experiments with drugs, enjoys sex with many partners, has little time for exercise, hates cooking and devours junk food. The time comes when, after serious reflection, she decides that she wants to have a baby. She's well aware that her lifestyle will have to change. She consults a close friend, a doctor, who advises modification before giving up birth control. Feeling the obligation to nurture *her future child*, she astounds those who know her with an overnight reformation.

This woman is admirable in assessing the obligations of a mother-to-be, and adjusting her conduct in their light. Those obligations aren't directed towards a specific individual. They are aimed at a being who satisfies a description 'the child I shall conceive'. Moreover, given the ways in which the timing of conception is sensitive to factors of maternal health, her action in giving up her old habits causes conception to occur earlier than it otherwise would. The child she bears is a different human being from the one who would have emerged from a conception and pregnancy under the old regime. Hence, if the latter pregnancy would have resulted in the birth of a developmentally damaged child, *that* child would not have been harmed by her continuation of 'partying-as-usual'. For, if the woman had reformed earlier, the damaged child would not have existed.

Or consider a related instance. A couple, at risk for having a child with some horrific genetic condition (Lesch-Nyhan or Sanfilippo syndrome), opts for in vitro fertilization (IVF). Several zygotes are produced, and tests reveal that some carry the awful genotype and others don't. Were they to choose one of the affected embryos for implantation, they would not have harmed the resulting child. For if they hadn't made that decision, the child would not have existed. Yet that perverse selection is surely a dereliction of pre-parental duty.

There's an obvious alternative vocabulary for describing cases of this sort. Forget harm. People take on—or are thrust into—roles. Those roles come with obligations. Failing to discharge the obligations is morally wrong. Whether or not that causes harm to some particular individual.

In light of cases in which the causal connections that generate the non-identity problem expose the need for thinking in terms of role obligations, rather than seeing the avoidance of harm as giving rise to the obligation, it seems preferable to approach stewardship in similar terms. To be sure, delinquency in discharging the duties of a role often causes harm to others. Not always, though.

Does this response simply pass the buck? What justifies assigning particular obligations to those assuming a role? I answer: in delineating roles we aim to make human lives go better than they otherwise would. That response depends on a conjecture: an ideal deliberation (of the kind envisaged in my moral methodology, and its extended version in ethical methodology) would endorse that aim, and, on that basis, the assignment of obligations to the roles of mother, parent, and steward. Obligations that apply even when nobody is harmed, but when failure to discharge the obligations would diminish the quality of some of the resulting human lives.

But who are the people who are to be represented in this deliberation? Some are people who already exist, but important stakeholders—the child, the visitors, the members of future generations—whose lives will be affected by the decisions taken now, remain unspecified. Does that matter?

We know nothing substantive about them, or their needs, or their wishes, or their psychology, except what can be inferred from the fact that they are human. What we can specify are features of their environments. The mother-to-be knows that the baby will grow in a uterine environment in which the biochemistry will be affected by her decision, whether she is abstinent or continues her partying habits; the steward knows that the visitors' pleasure will be enhanced or diminished by the state of the garden; anyone who thinks carefully about global heating recognizes the environmental shocks resulting from increased global mean temperatures. If we are focused on the quality of the lives, if representing the

unspecified stakeholders depends on assessing the quality of lives lived in different environments, the deliberation has enough information to be carried through.

There is an alternative line of defense available in the case of most concern to us. In considering how best to cope with global heating, we can start with people who are already here. Younger members of the contemporary human population will exist in a warmer world than the one we know. How much warmer that world will be depends on the decisions we now make. Some of them can stand in for the human future, presenting their own perspectives as representative.

But *how* representative? None of those who speak now can speak for people who will live a century hence. A discussion taking them as representatives of the future will be biased towards the short term. True enough. Yet these deliberators will be able to recognize their own future predicament, living in a world in which they have ties to descendants who already exist and whose lives matter to them. Just as we expect today's elderly to be affected by the fates of their younger contemporaries, seeing the quality of their own lives as dependent on how well other lives, lives that will long outlast their own, will go, so too for the younger participants in today's deliberations. They have a stake in the more distant future. Moreover, those who will play the part of the young when they are elderly will have a stake in a yet more distant future. And so it goes. We see here the first implications of the Wittgensteinian image of my epigraph—an image to which we shall return shortly.

Like us, future people, whoever they turn out to be, will require food and water, shelter and protection from predation. If some of the structure of current societies has been preserved, they will want to be educated, to maintain their health, to enter into close relationships with others, to form their own plans of life. If that structure has been weakened or destroyed, their representatives in our deliberations will reasonably attribute to them a desire for it to be restored.

Representatives of future generations must convey to the other deliberators what it would be like to live in each of the habitats resulting from the options under consideration. Or, more exactly, they should assess the chances of the various kinds of changes potentially resulting from a projected policy, and how such changes would bear on the lives of the people who would inhabit that kind of world. Their task is to facilitate a particular style of sympathetic understanding. Not 'feeling the others' pain', but the second species of sympathy Adam Smith identified: becoming aware of how you would be affected in another's situation. Successful representation of these unknown future people is a succession of vivid depictions of the environments likely to come about if the policies under consideration are pursued. Mutual engagement occurs to the extent the deliberators representing present constituencies are able to comprehend, cognitively and affectively, what it would be like to live under these conditions.

At this point the entanglement with epistemic issues should be evident. Appropriate representation requires absorbing the many kinds of information presented in IPCC reports. What is the best estimate for the global mean temperature, given a particular scenario? What's the best estimate of the range of temperatures to be experienced by those living in a particular region? What is the probability of extreme weather events (heatwaves, for example)? But far more is needed. Attention to mean values, and even to distributions, of important climatic variables is only the beginning. Deliberators will need to understand how those values and distributions contribute to events that cause radical disruptions of human lives. And more besides. How the disruptive events interact with one another—how the prolonged drought produces migration, and outbreaks of disease, and breakdown of

hygiene, and interactions with neighbors who are already struggling and . . . . Forecasters are always in danger of selling the difficulties of a changed climate short. (Wallace-Wells 2019 provides an outstanding survey of challenges that are often ignored, but, as I suspect he would admit, his own inventory is far from complete.)

A realistic vision of future human life so far outruns the information climate scientists can reliably offer as to defeat any attempt at representation that is vivid, comprehensive, and justified. Hence, a critic might charge, the approach I have recommended is doomed by the epistemic demands it makes. I concede: my approach is the worst we could try—except for all the others. For any attempt to sort out our obligations to future people will find itself in the same boat. Whatever views anyone has about the pursuit of this moral inquiry, they will have to be informed by some justified vision. Woeful ignorance is the common condition for any exploration of our stewardship obligations.

Yet that ignorance applies unevenly across the human future. And that offers the way to a more nuanced treatment of the issue.

* * *

My discussion so far has, for the most part, lumped future people together, whether they live 10 or 100 or 100,000 years from now. Can I do better than that? Let's fantasize. Suppose we had sufficient information about the future to construct realistic visions of the environments our descendants will experience, at 100-year intervals from now until the best estimate of our species' extinction, for all of the policies we envision, from business as usual to the most radical proposals for a transition to a sustainable planet. The ideal deliberation could then include a very large number of representatives of different parts of the future, and weigh their individual assessments of the rival merits of the full spectrum of options. If it were to live up to the three suggested conditions, it would yield a (revisable) solution for the long term.

Plainly we can't manage anything remotely like that. So we're forced to choose the moments (or periods) in the future on which to focus. How to select? A maxim sometimes beloved of social scientists suggests an answer: look for your lost keys under the lamppost, because that's where the light is. Concentrate, then, on the future times for which current predictions about the future are best justified.

Climate modeling is difficult because of the number of potentially relevant factors. Modelers make different—reasonable—selections, run their simulations, and draw conclusions about some set of focal variables. They may be concerned with the pH of an ocean, the mean temperature of a region, or the deviation in some persistent airflow—or any number of other things. As is well known, their expert practices standardly generate a collection of estimates. A plausible principle might take the reliability of the mean to be inversely correlated with the size of the range. Hence, in deciding on a time at which to try to predict the future, we might seek the point at which the spread among predictions is narrowest.

Which time should that be? Climate modelers are the appropriate judges, but my own (outsider's) reading of work in this area suggests a tentative figure: half a century from now. Shorter-term predictions tend to scatter more broadly, scenarios converge 40 to 60 years hence, and then diverge ever more radically. For present purposes, then, and awaiting expert refinement, I'll take 50 years to be the appropriate target.

The representatives of the future, on this account, should be people who are skilled at interpreting how the mean predicted values, half a century from now, will affect the

environments experienced by our successors who will live then. The task of the deliberation will be to understand their predicament, to consider the situations of members of the three other constituencies, and to find some way of distributing burdens all can tolerate. To follow that policy is the best—the *morally* best—we can do.

Or is it? Proceeding in this way seems to violate one of the three conditions on ideal deliberation: it fails to include representatives of *all* those who will feel the impact of whatever is decided. Moreover, failures of this kind are conspicuous in our past, retarding and often blocking changes we regard as paradigms of moral progress. The abolition of slavery, the expansion of opportunities for woman, the appreciation of same-sex love all had to overcome the exclusion of the voices of oppressed people, the slaves, the women, the gays and lesbians. Haven't I made the same mistake in omitting two classes of stakeholders: those who will live in the near-term future, and, more importantly, those who will inhabit our planet a century, a millennium, or a 100,00 years hence?

The short-termers are easy to accommodate. If an environment 50 years from now is tolerable to those who exist then, the less problematic environments of the intervening years ought to satisfy the people who live through them. The issue, then, is whether the agreed-on policy asks more of them than it does of the world's current inhabitants (represented already in the deliberation)—as it would, for instance, if the agreed-on policy deferred cutting emissions for a decade or so, and then insisted on a rapid change to forgoing all non-renewable sources of energy. The obvious solution (as I have already hinted) is to include young deliberators, a large part of whose lives will lie in the coming decades, and to require them to think not only for themselves but for their children.

The serious problem is myopia. The deliberation can be viewed as identifying the role obligations for a short-sighted steward, someone who diligently maintains a garden—or the whole planet—for the immediate future, but does so in ways that create, or permit, long-term trouble and even disasters. My proposal needs refinement.

Even the myopic are sometimes able to recognize an impending threat, represented in a sudden blurry image. So too in the present instance. My previous remarks about the wide scattering of predictions in the distant future are correct—but they don't rule out the possibility of identifying features of our planet that are vulnerable to change (often changing monotonically, year by year) whose state is important to human well-being. We can observe the melting of ice, on mountain tops and in polar ice-sheets, and can trace the consequences for water flows to heavily populated areas and for sea levels. We can know now that, if these processes continue, there will be severely damaging effects on human life some centuries hence, well beyond my predictive horizon of 50 years. A feature of the Earth whose presence is required if some environment for our descendants is not to become severely challenging, and which is currently being eroded as a result of human activities, is a *necessary retention*, even though the decay may be slow and the effects on human life not felt for thousands, tens of thousands, or hundreds of thousands of years.

To cope with myopia, stewards have to take account of necessary retentions. That means including in the deliberation representatives of the longer-term future (the indefinitely extending period beyond my 50-year horizon) who can use contemporary knowledge about our changing planet to identify necessary retentions, explain the timescales on which the monotonic processes will operate, and interpret the consequences for future populations. This is one way to accommodate the legitimate demands of people in the distant future that they, too, should have a voice in framing what should be done.

It is not, however, nor should it be, the only way.

* * *

Acknowledging ignorance is important. Whatever desire (or obligation) we have to plan for the long-term future, we can't do so by guessing. To adapt a famous Wittgensteinian dictum: What we cannot speak about, we must pass over in silence. Apart from identifying necessary retentions, responsible planning is restricted by our limited powers of prediction.

Yet the fact that *we* cannot now predict with confidence beyond a limited horizon (50 years in my tentative suggestion) doesn't entail that the interests of people who will come to be beyond that horizon are doomed to be slighted. For the deliberative process I have envisaged is not a one-shot affair. It can be repeated. Later repetitions bring what were previously inaccessible regions of the future into view. Imagine, then, a sequence of deliberations, say at 10 year intervals, not necessarily taking half a century as their target, but, because of the progress of climate science, forecasting with respect to an expanded interval. Each discussion builds on, and adjusts, the policies accepted by its predecessors. We reach the distant future by a series of stages at which we plan responsibly for the short term (and perhaps by recognizing an ever more extensive set of necessary retentions). In the spirit of the Wittgensteinian analogy that serves as my epigraph, our planning for the long term is morally justified not because it is achieved all at once in some (impossible) revelation, but because it consists in a set of overlapping morally justified policies.

A responsible steward cannot simply plan for the immediate years to come. Where long-term threats can be foreseen, steps must be taken to ward them off. Beyond that, the steward must prepare so that future stewards can do the same conscientious job.

Recognizing this has implications for deliberations about climate policy. The task is not to arrive at a final strategy, to be graven in stone and pursued come what may. Rather, in the spirit of the image William James invokes at the end of 'The Will to Believe', of travelers on an icy Alpine pass in a blinding snowstorm, we have to strike out in whatever direction appears most promising and modify our course in light of what we learn as we go. Not only must we set up conditions under which future deliberators can adjust plans and policies to accommodate the lessons of their richer experience; we should also strive to make their deliberations go better than ours.

This means, first and most obviously, that the conditions required for large groups of experts to find out as much as possible about how the planet is changing must be maintained—if not improved. Moreover, that technologies required for implementing our plans must be supported and distributed as widely as possible. It seems likely to me that these apparently modest requirements will issue imperatives for the modification of global capitalism as we know it, beginning, for example, with reform of the idea that the distribution of beneficial breakthroughs must always be constrained by obeisance to private profit.

The latter implications are even more evident if the goal of improving deliberation is taken with the seriousness it deserves. So far, the most evident concern about the moral methodology I have adapted to the case of climate policy has been studiously avoided. As many people have reminded me, the conditions on ideal deliberation are too far from the actual world.[5] Perhaps if the entire human population consisted of unusually thoughtful,

[5] Here, I take up the concern about whether my deliberative ideal can allow any realistic implementation.

other-directed, and public-spirited people, something akin to the discussions I envisage might take place. Too many recent events reveal a very different situation: rancorous, selfish, thoughtless, ignorant people won't even try to reach agreement with those they perceive as enemies (and often as evil). If I am right in treating the methodology of moral inquiry as I do, then, with respect to the major moral issues now confronting us, we are doomed to fall alarmingly short.

Existing discussions of climate policy, held regularly for over three decades, reveal the enormous distance between a genuine moral inquiry and how plans for the future are actually negotiated. Apart from the obvious facts that important parties (the poor and powerless all over the globe) go unrepresented and that the degree of mutual engagement is pitiful, these meetings proceed by assuming cooperation is a temporary expedient, to be jettisoned when the crisis is past. Hence national policymakers operate under the shadow of the future, envisaging a world in which dog-eat-dog capitalism resumes its proper economic dominance and gestures towards reducing inequalities no longer need to be made. The obvious incentive is thus for each nation to jockey for a favorable position once competition resumes in full force, and to reduce its own relative share of the burdens. As a number of writers have cogently argued, it is hard to see how such 'deliberations' can sustain a transition to a sustainable environment (Klein 2014).

One obligation of stewardship is to enable our successors to do better than we can currently manage. It can be partly discharged by agreeing on explicit norms for deliberation that demand attempts at mutual understanding and mutual engagement, using existing experiments in deliberative democracy as sources of inspiration (Fishkin 2009; 2018). These ventures can be aided by changes in education aimed at creating citizens who are more skilled at moral deliberation than their contemporary counterparts: people whose responsiveness to others has been appropriately amplified (Kitcher 2021b: chs. 4–5).

If global emissions are not reduced drastically during the next few decades, there are very high probabilities of environments in future decades (even within this century) that will make the lives of more than a billion people extremely difficult. There are significant probabilities that the consequences of their predicaments, because of their understandable efforts to move elsewhere, will affect the entire human population: that pandemics and wars will ensue.[6] The chances of the global mean temperature increasing by more than 3°C above pre-industrial readings can no longer be dismissed as tiny. Admirable efforts to reduce individual carbon footprints by a minority of people in a minority of the world's nations will make no difference to these estimates. Without a pan-human alliance to combat a common enemy—the greenhouse gases we have emitted, first unwittingly, more recently in full knowledge of the danger—the chances of large-scale threats to human existence will only increase. Unless technological developments resolve the problem, most likely through making it possible to remove carbon from the air and to store it safely,[7] thus returning the atmosphere to a state of which we have prior experience, that alliance must be renewed throughout the indefinite future.

[6] Several readers of Kitcher and Keller (2017) who were initially skeptical about the realism of our opening apocalyptic scenario have since written to me to say that it no longer seems far-fetched.

[7] I allude here to a distinction between two technological 'fixes' that are importantly different. Subtracting carbon returns us to a world we have already experienced, and know we can live in; adding ingredients to the atmosphere (sulfur particles, for example) launches us into an uncharted space. See Kitcher and Keller (2017: 98–103).

Creating and maintaining that alliance requires responsiveness on a scale the human population has never previously achieved. Representatives of the world's nations have been holding periodic meetings ever since 1990, and the outcomes have been depressingly similar. Inadequate targets are set. The efforts to meet them fall far short. There are apologetic mutterings and promises to do better. The probability of dire consequences increases. New targets—usually still more inadequate—are set. The cycle repeats.

Without clear commitments to address the needs of all four constituencies, the frustrating pattern will continue, and the perils will loom ever larger. Our species urgently needs to address the problem my ideal of deliberation diagnoses. The meetings to which I have referred are inclusive (all nations attend). They are reasonably well-informed (perhaps the IPCC is too conservative in its assessments of the dangers, but its reports are at least as good as those guiding other policy discussions). The trouble is the failure to engage. Without seeing the world through the eyes of the precarious poor and the delayed developers, feeling it through the skins of those who will outlive almost all of the participants, and recognizing the concerns of the cultural conservators, there is no hope for a program that will both address the predicament and prove politically viable when the meeting is over and the delegates return to their homes. Participating on conditions of profound inequality, nations jockey for future economic advantages, and the worldwide agreements we need are doomed.

In wartime, people have sometimes managed to do better, even to cooperate with allies whom they disliked. Those human adjustments point the direction. But we must go further. Sympathetic identification with the full range of human forms of life must go deep enough to enable the alliance to stay firm, and to be renewed, generation after generation. The competitive capitalism spawned in the 1980s must give way to an appreciation of a point appreciated by insightful economists from Adam Smith to Alvin Roth: markets must be designed to do the things we identify as most important. In a world where our species faces a serious threat of short-term extinction, it is hardly a consolation to be informed that we have been enabled to buy a diverse basket of consumer goods at very low prices.

We shall need to achieve what I have elsewhere dubbed the 'Deweyan Society' (Kitcher 2021b, especially ch. 10), in which socio-economic relations within and among countries are radically reformed. Without a technological *deus ex machina* there is no alternative.

* * *

I conclude with some brief remarks about the other long-term challenges noted at the beginning (although perhaps the last few paragraphs indicate why I see the climate issue as casting a long shadow over all of them). In each instance, I suggest, the basic problem is the same: we shall need a pan-human program for agreeing on common policy, an alliance that must be renewed into the indefinite future. To attain that, we'll require just the amplification of human responsiveness necessary in the climate case, and the same socio-economic reforms at which I have gestured. These alternatives are different only in the variations underlying the epistemic horizons. The segmentation of the future into periods depends on how far, with respect to particular threats, reliable predictions can be made. Inclusive, informed, engaged deliberation crafts policy for that part of the future where conditions can be reasonably foreseen, leaving to our successors who live at the terminal point to determine it, in the same or in a refined fashion, for the subsequent interval. As before, any

deliberation must attend to ensuring that the conditions for proper deliberation will be available at the next stage.

The climate problem, then, provides something of a blueprint for coping with other long-term challenges. Perhaps this essay also explains why I view it as dwarfing all the rest.

References

Fishkin, J. (2009), *When the People Speak* (Oxford University Press).
Fishkin, J. (2018), *Democracy When the People Are Thinking* (Oxford University Press).
Kitcher, P. (2011), *The Ethical Project* (Harvard University Press).
Kitcher, P. (2015), 'Experimental Animals', in *Philosophy and Public Affairs* 43: 287–311.
Kitcher, P. (2018), 'Governing Darwin's World', in P. Adamson and G. F. Edwards (eds.), *Animals: Historical Perspectives* (Oxford University Press), 269–92.
Kitcher, P. (2021a), *Moral Progress* (Oxford University Press).
Kitcher, P. (2021b), *The Main Enterprise of the World: Rethinking Education* (Oxford University Press).
Kitcher, P. and Keller, E. (2017), *The Seasons Alter* (Norton/Liveright).
Klein, N. (2014), *This Changes Everything* (Simon & Schuster).
Parfit, D. (1984), *Reasons and Persons* (Oxford University Press).
Wallace-Wells, D. (2019), *The Uninhabitable Earth* (Penguin).

13

Shaping Humanity's Longterm Trajectory

Toby Ord

1 Introduction

Humanity may have a very long time ahead. Our species has already survived about 300,000 years, and the typical species lifespan is roughly 1 million years (Barnosky 2011; Galway-Witham and Stringer 2018). Some species, such as the horseshoe crab and the nautilus, have survived unaltered for hundreds of millions of years. One kind of upper bound on humanity's longevity comes from the Earth itself, which will remain habitable for roughly 1 billion more years. But even that is not a hard limit: if our descendants migrate to other stellar systems with newer (or longer-lived) suns, they could survive for at least a thousand times as long again (Adams and Laughlin 1997; 1999). So if we play our cards right, humanity could have a flourishing civilisation that lasts for a time that is almost beyond comprehension.

Longtermism is a moral outlook that takes this vast future seriously (Ord 2020: 43–49; Greaves and MacAskill 2021; MacAskill 2022). It considers how the possibility of making lasting alterations to humanity's longterm future might inform the actions we should take now.

One clear way humanity's longterm future might be altered is via *existential risks*—risks of irrevocably destroying humanity's longterm potential, such as through extinction or an unrecoverable collapse of civilisation (see Bostrom 2013; Ord 2020). Examples include the risks of asteroid impacts, supervolcanic eruptions, nuclear war, engineered pandemics, and climate change. Existential catastrophes provide a clear example of how our actions today could have effects that don't simply wash out over time. This is because:

1. they could occur in our own time,
2. if they did occur, they rule out the chance of a subsequent recovery, yet
3. they could be prevented by concerted human action.

This means that an action to increase or decrease existential risk could be targeted at a near-future event yet have an expected value that scales in proportion to the entire size of humanity's longterm potential.

But avoiding existential risk isn't the only way our actions might shape the longterm future. For example, we may be able to temporarily—or even permanently—speed up progress: allowing people earlier access to the prosperity, technology, or social progress that would have come in later years. And if there is a chance that humanity's values get locked

Toby Ord, *Shaping Humanity's Longterm Trajectory* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Toby Ord 2025. DOI: 10.1093/9780191979972.003.0013

in at some stage in our development, then actions to improve our values now may lead to us being guided by better values for millions of years (see Ord 2020: 153–8; MacAskill 2022: 78–9). Or more generally, if there are multiple longterm equilibria for our society, early actions may be able to change which one we end up in, and so have lasting effects.

2 Longterm trajectories

Consider the trajectory of human history up to the present day, and then imagine some of the ways it could continue far beyond, into the distant future. We might have a short future of unconstrained technological progress ending in an extinction event in the next century. Or we might have a much longer future where we live modestly, in harmony with nature, for a typical species lifetime of a million years. Or perhaps we might have a future of epic proportion: spanning trillions of years and billions of worlds.

The purpose of this essay is to find a way of formally representing such trajectories that helps us understand and compare the many ways we might have lasting effects upon them. Doing so involves a balancing act between abstracting away enough details to make the representation workable, while leaving enough to provide the power to make fruitful comparisons.

There are many ways we might represent such trajectories. For example, we could characterise humanity's situation at any one time as a point in a multidimensional space, with dimensions that represent all the key variables for our civilisation, such as our technological capacity, our co-ordination, and our wisdom. We could then imagine the trajectories as the paths our civilisation might trace out in this space as its combination of attributes changes.[1] Such an approach has been suggested by Beckstead (2013: 69–73) and Bostrom (2013).

Or we might focus on a single dimension (such as our technological capacity) and consider a graph of how this could rise or fall over time. Bostrom (2013) and Baum et al. (2019) adopt such an approach, using it as a way of illustrating and teasing out different *qualitative* scenarios, such as stagnation, collapse with the possibility of later recovery, or an ephemeral success followed by premature extinction.

I want to explore a somewhat different approach. Like Bostrom, and Baum et al., I will focus on a single representative dimension of humanity and how that evolves over time. But I use a special dimension, chosen to allow quantitative evaluations and comparisons of trajectories.

My primary focus is the value achieved by humanity over time. On the vertical axis is the *instantaneous* value of humanity at any moment in time. This means that the total area under the curve represents the cumulative value of humanity over all time. For example, we could think of the vertical axis in terms of value per year, the horizontal axis in terms of years, and thus the area under the curve in terms of value.

Exactly what this represents depends on our theory of value. As a simple example, if our theory of value were classical utilitarianism, then the height of the curve would be the total amount of happiness minus suffering occurring at that time, and the total area under the curve would be the cumulative amount of happiness minus suffering created in the whole of human history.

[1] An interesting choice for such an approach is whether to distinguish trajectories that trace the same path through state space but proceed at different speeds or linger at different points on the path.

But the approach is flexible enough to encompass a wide range of conceptions of value. It could have a richer story of what contributes to an individual's wellbeing; it could ascribe intrinsic disvalue to inequality in wellbeing (at a given time); or it could ascribe value to things other than wellbeing, such as art, knowledge, achievements, or the environment.

The main constraints are:

1. Value is something that is at least roughly quantitative, such that it makes sense to add it up.
2. Value is separable across time, such that we can 'date' contributions to intrinsic value to a particular time and find the total value by the integral (or sum) across these times.[2]

This is compatible with a range of theories of population ethics including: the total view, the critical level view, a version of the average view that values the integral of the average wellbeing at each time, some person-affecting views, and theories that have diminishing marginal value for additional population at a given time. But it is incompatible with views that are not time-separable (see Broome 2004), such as the average wellbeing of all people who will ever live, or a view with diminishing marginal value on the total number of people who will ever live

The chief reason for adopting this framework is that by focusing on value as the key dimension, it allows us to evaluate and compare different trajectories—and even different *changes* we might make to a trajectory. The hope is that this will inform us about how we should strive to change humanity's future. But an important limitation is that it will only produce considerations related to value. If there are important parts of morality that are unrelated to value (such as personal freedoms or unbreakable rules) the framework will remain silent about those. It will just tell us how much value is at stake and leave these other considerations up to us.

To see what this framework allows us to do, let's start with an illustrative trajectory for humanity (see Figure 13.1).

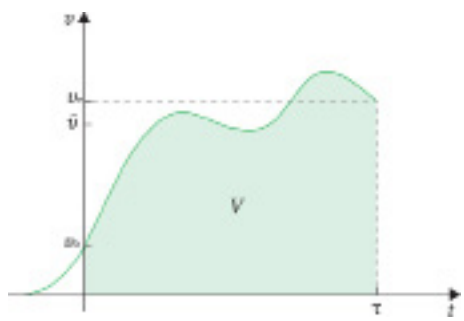


Figure 13.1 An illustrative longterm trajectory of humanity. On the vertical axis is the instantaneous value of humanity and on the horizontal axis, time. This means that the shaded area under the curve represents the cumulative value of humanity's future.

[2] In my presentation of the framework, I'll treat things as if value is ascribed to particular instants and so is integrated over time. But a lot of what I say would be the same if it were instead ascribed to longer periods and was summed over time. The main issues are that we may no longer be able to apply the concepts of a continuous or smooth curve and that it may be messy if an intervention advances or delays progress by some fraction of a time period. Neither are big issues if the periods were very short (e.g. minutes) but it may become a substantial issue if the relevant periods are generations or centuries.

This trajectory begins with rapid and escalating growth in the instantaneous value of humanity that slows towards a peak, then gradually declines before rising to an even higher peak. After a final period of decline, humanity's trajectory abruptly ends (perhaps in catastrophe; perhaps after achieving all it could). I have labelled some important features that apply to any trajectory:

$t \equiv$ time (in years)
$t = 0 \equiv$ the present time
$\tau \equiv$ the endpoint of humanity's future = the duration of humanity's future
$v \equiv$ the instantaneous value of humanity
$v(t) \equiv$ the instantaneous value of humanity at a given time
(the function $v(\cdot)$ itself can be used as a name for the whole trajectory)
$v_0 \equiv$ the instantaneous value of humanity at the present time = $v(0)$
$v_\tau \equiv$ the instantaneous value of humanity at its final time = $v(\tau)$
$\overline{v} \equiv$ the average instantaneous value of humanity from time 0 to τ

$$V = \text{the total value of the future} = V = \int_0^\tau v(t).dt = \overline{v}\,\tau$$

As we have seen, the duration of our future could be truly vast. Most species in our position could look forward to about 10,000 more centuries, and with our unique capabilities we have the potential to last for billions of centuries. So one way in which the longterm future could be vastly more valuable than our current century is through its duration, τ. The assumptions embedded in this framework imply that the value of our future scales linearly with this duration—other things being equal, a future of a thousand times the duration is a thousand times better.

A second way our future could be vastly more valuable than the present is by being much more valuable at any given time. This could be true if we are able to build fairer and more just societies with much less of what is bad in life. It could also come from us each having much more of what is good in life. Since our peak experiences far outshine the average, there is room for each of our lives to be vastly better if only we could spend longer at those heights.

And civilisation of the future may also be much more valuable through being so much larger than what we have today. There are more than 100 billion planets in the Milky Way, so the scale of our future civilisation could be vastly greater than it is now (Cassan et al. 2012; Ord 2021). And this may matter a great deal. Or it may not. It isn't clear how to evaluate such increases in the scale of humanity at a time, and there is much disagreement. But this is a plausible way that the value of humanity at future times may be much much greater than it is today.

To see how much more valuable than our own time the entire future could be, let's shade in the value of our current century on the trajectory we considered earlier (see Figure 13.2):

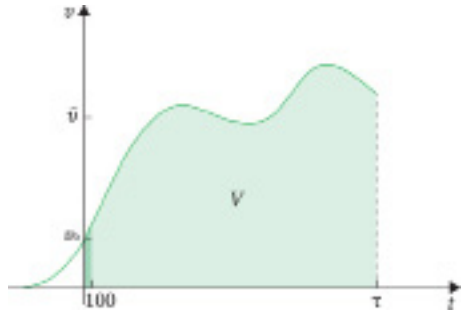


Figure 13.2 Comparing the value of the next 100 years (dark shading) to that of the entire future (light shading). Because the length, height, and shape of the future trajectory are extremely uncertain, this diagram shows the key parts that need to be compared but is unlikely to represent them to scale.

Now consider the ratio of the value of our present century to the value of the entire future. The current instantaneous value of humanity is v_0 units of value per year, and so (assuming the average value across the century isn't radically different from the value right now) the value of our century is roughly v_0 times 100. The value of the entire future is V which can also be expressed as $\bar{v}\tau$. So the ratio between these is roughly:

$$v_0 100 : \bar{v}\tau$$

We could also think of this in terms of the question:

How much more valuable is the entire future than the current century?

Rearranging the expression above, we get a multiplier comprised of two factors:

$$\frac{\bar{v}}{v} \cdot \frac{\tau}{100}$$

The first factor is how much better the average century is compared to our own, and the second is how many more centuries there are. These are two quite different ways the future could be much more valuable than the nearer term. And each of these factors might be truly vast—quite possibly a billion or more.

Put another way, the sheer duration of the future is more than enough to get longtermism off the ground. And the sheer scale (or quality) of the future at any one time could be too. Either one alone could suffice. And yet because they are multiplicative, there is also a very real possibility that the true scale of future value could be more than a billion *billion* times that of the next century.

If we were just trying to make the case for a greater focus on the longterm effects of our actions, this observation would be excessive. The most robust and persuasive case for focusing on the long term would be to focus on the most widely accepted part—its

duration—and leave the heightened value at a time out of it. But as my aim here is not to argue for longtermism, but to develop a technique for comparing different ways of shaping the longterm future, we need to keep track of both dimensions. Later, we will see that some interventions scale in value with just one or the other of these factors, making comparisons between these interventions depend crucially on the relative sizes of the factors.

3 Aims

It is worth taking a moment to explain the aims of this framework, and how it fits with work in economics and ethics.

The chief aim is to help us understand and compare different ways of altering the longterm trajectory of humanity. The framework's focus is on the cumulative value of humanity over its entire duration, in particular, on how changes to a range of key parameters of that trajectory affect the total value achieved. While the changes to the parameters may be relatively small (e.g. reaching each point in our development one year earlier, or lowering extinction risk by 1 part in 1,000), their effects will often be vast, as they may be felt over the entire future.

Achieving this aim requires the framework to be quantitative, so that we can use mathematical techniques to analyse what is happening. But many of the results will be qualitative. For example, showing that one kind of intervention scales in value with the duration of our future while another one does not. Understanding the ways that different interventions scale with the shape of the future helps us see when one intervention is really of a different kind to another, and so will help us develop useful categories of longtermist interventions.

I will illustrate many of the ideas with diagrams. These are designed to clearly show what intervention is being considered and how its consequences unfold. Doing so almost *requires* that the diagrams are not drawn to scale. One reason is that we are primarily considering changes that are in some way small compared to what is displayed on the graph—changing things by 1 part in 1,000 wouldn't show up clearly on a scale diagram. But this is a smaller loss than you may think, since there is so much uncertainty in the scale and duration of our future anyway that one really can't hope to draw it to scale in any definitive way. But we do need to bear in mind that in a true-scale diagram, the entire duration of human history to this point in time may be a vanishingly narrow sliver at the start of the trajectory—brief in time compared to the aeons ahead and perhaps meagre in instantaneous value compared to what will be able to be achieved with our full maturity.

One useful point of comparison is with supply and demand curves in economics. These are quantitative curves but are often drawn in a highly stylised manner. They help make qualitative distinctions and develop our intuitive understanding of how various effects scale as a parameter is changed (see Figure 13.3).

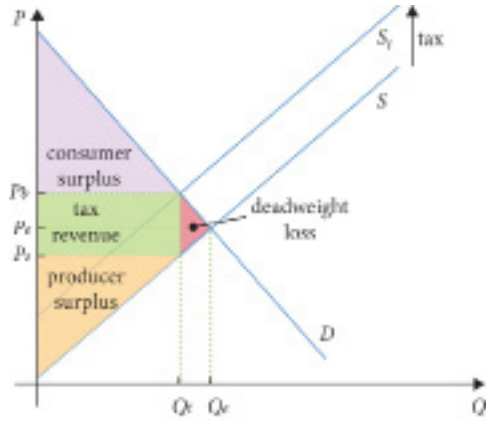


Figure 13.3 Supply and demand curves showing the effect of adding an excise tax. Like the diagrams I will introduce, this consists of studying what happens to the areas between certain curves as those curves are shifted or adjusted. And its key insights can be gleaned even when the curves are simplified and not drawn to scale.

The theory of longterm trajectories involves aspects of both ethics and economics, and could be studied within either discipline.

All the objects of study—value, time, choices, consequences—are at home within ethics. But the use of mathematical methods to study them is not so common there. Nor is the focus on value over such large scales, or on choices that shape the course of history.

In contrast, the wide scope of optimal planning over the whole future is familiar in economics, as are the mathematical methods used to analyse it (e.g. in the study of optimal growth or intergenerational equity). But the thing being optimised—value—is a little unusual. It is closest to utility, but can have several differences, most notably that it is assumed to be a ratio-scale quantity and to be time-separable. This is what allows us to treat it as a flow of instantaneous value over time and see total value as the integral of that flow (see John Broome (2004) for how this can be done in ways that should satisfy both the philosopher and the economist).

The other main point of departure from most economic analysis is that I use a zero rate of pure time preference.

4 Discounting

Economists frequently study flows of benefits and costs over time. They typically evaluate such flows using a technique called discounting. This means applying a mathematical function that reduces ('discounts') the relative importance of costs and benefits that occur further into the future. They do this for a variety of different reasons.

One key reason is that economists are usually considering monetary benefits or costs. Money doesn't translate directly into value, so its value may depend on when it is received

or paid. Money often needs to be discounted due to empirical facts about the interest rate, the growth rate, and the diminishing returns to wellbeing from having more money. Earlier monetary costs are magnified compared to later ones, because they mean you miss out on investing the money over the intervening years; later monetary benefits are typically worth less than earlier ones because the people who receive them are richer, so get less value from each dollar. These are good reasons for discounting monetary benefits and costs, but they won't apply here as we are directly concerned with evaluating streams of value (or utility) itself.

Another reason offered for discounting the future is that people have a brute preference for value to come sooner rather than later ('pure time preference'). For example, a person may simply prefer one treat now to two later. It is not actually so clear whether people do have such preferences, since it is difficult to experimentally distinguish between people having a considered preference for immediate benefits versus suffering weakness of will where they act in a short-termist way against their own better judgment.

But regardless of whether people have such preferences with respect to their own lives, experimental evidence suggests they don't have them regarding benefits to other people, such as one's children or future generations. Faced with such choices, people are roughly indifferent to the timing of the benefit or cost (Frederick 2003). Since these are the kinds of choices under discussion when considering the longterm future of humanity, and because there are also strong moral arguments for treating people equally whenever they happen to live, I will follow the consensus of moral philosophers (and some eminent economists such as Pigou (1920: 25), Ramsey (1928: 543), and Harrod (1948: 40)) in rejecting this kind of discounting.

A third reason for discounting is based on risk. An individual might rationally discount the value of benefits they would receive in their eighties or nineties on account of the reduced probability that they are alive to receive them. Similarly, the chance that humanity is still in existence will monotonically decline over time.

This could certainly provide reasons to discount future benefits (Ng 2005). But it doesn't apply to the methods used in this chapter. Here I'm focusing on the *ex post* value produced by an intervention in a particular outcome. Risk could then enter the picture by considering prospects over these outcomes. If this includes uncertainty over the length of the trajectory, that will produce an effect similar to discounting, though one that isn't baked in via an exogenous hazard rate (such as Stern 2007). Instead, it would be more flexible and expressive, allowing for variable hazard rates and for the choices we make to affect those rates. But this deeper treatment obviates any need to discount the individual *ex post* outcomes we are considering in this chapter on grounds of risk.

Finally, some economists (e.g. Koopmans 1960) have suggested that, whether or not there is a good reason, we simply *must* discount future value because otherwise the sum of value over time becomes infinite, causing intractable mathematical problems when evaluating or comparing different choices.

There are real challenges here, and they affect all approaches unless certain restrictions are made. Even exponential discounting fails to resolve the problem unless there are restrictions to prevent the possibility of instantaneous value growing exponentially over time. In the present work I avoid the problem by restricting the scope to trajectories that are finite

in duration.[3] This is not much of a restriction in practice. Instead of considering infinite flows of value, I just consider finite flows with longer and longer durations. The duration of humanity's future appears as a parameter, τ, and we can ask questions such as:

- Which intervention is superior in the limit of large τ?
- Where is the cut-off for τ beyond which intervention 1 has a better effect than intervention 2?
- How does the value of this intervention scale with τ?

This framework thus has an unusually low amount of discounting compared to most related work in economics. And this will cause some challenges.

When there is ample discounting of the future, one doesn't have to worry much about how to model the very longterm impacts of the intervention you are studying. The impacts of the longest-run effects are so dampened by the discounting that their entire value from 100 years out to eternity is usually bounded and small.

This is convenient. For without discounting, the comparisons between different interventions could be mainly driven by the assumptions about their longterm effects—assumptions that are quite unreliable if they were not the focus of the economic analysis. Heavy use of discounting (especially pure time preference) avoids this.

But it can also throw the baby out with the bathwater. What if the main effects of our everyday policies really are their longterm effects? Or what if we want to study the set of interventions targeted towards changing the long-run future? If so, then heavy use of discounting, while convenient, would be inappropriate. These longterm effects are the very object of our study. They need to be illuminated and clarified, rather than swept aside. Indeed, we could even go so far as to say that one of the purposes of this framework is to allow us to analyse and compare very long-reaching effects of our actions without needing to discount them.

5 Idealised changes to the trajectory

When analysing lasting changes to humanity's longterm trajectory, we are often interested in marginal changes—that is, relatively small adjustments to our trajectory from the present time, $t = 0$, onwards.

Why marginal changes? Since the overall trajectory could be so much longer than everyday timescales and the average instantaneous value so much greater than what we see today, even changes that are very large in today's terms could be relatively small compared to the whole future. Relatively small changes to the longterm future may thus be the best we can do. And yet they may still matter a great deal in absolute terms, as their value could accumulate over deep time. Marginal changes are also easier to analyse, allowing us

[3] When extending this work to deal with uncertainty, the challenges would return due to the possibility of infinite expected value even when the duration is certain to be finite (e.g. consider a survival function with a tail that decays as $1/t$). In such cases I'm optimistic about resolving some of the challenges by assigning these hyperreal expected values, in a manner related to Pivato (2008) and Bostrom (2011).

to imagine everything else being roughly the same despite the change. Of course, marginal changes are not the whole story, but they are a key component and a good place to start.

Mathematically, we will be analysing a marginal change to some key parameter of the trajectory of humanity, adjusting it by some small amount. We will use the symbol δ (as in δt) to refer to these marginal changes in a parameter (reserving the capital letter Δ for large changes, to be studied at a later date). Each marginal change will transform the trajectory $v(\cdot)$ from 0 onwards. We shall represent the transformed trajectory as $v^*(\cdot)$ and its endpoint as τ^*. We can then examine the effects of this change upon the value of the future (the integral of the trajectory from 0 to τ^*).

In general, evaluating the effects of even a small change to the shape of the future trajectory would require detailed knowledge of the shape of the default trajectory—the course history was going to take before we intervened. Given our ignorance of this shape, that would make such evaluations extremely difficult. But there is a family of idealised changes to the trajectory that can be evaluated without detailed knowledge of the default shape. These idealised changes depend only on a few of the key parameters we've seen earlier, such as the duration of the future, τ. While these parameters are *also* unknown, the quantities we care about will now be expressible as simple functions of these unknowns.

The kinds of idealised change we will explore are:

- Advancements
- Speed-ups
- Gains
- Enhancements

In reality, the interventions open to us will be much messier than any of these idealised versions. But due to their simplicity and tractability, the idealised changes will provide a useful starting framework for analysis.

6 Advancements

One way to change the future is to advance progress. If we think the future is likely to be better than the present, we could try to reach those higher levels of instantaneous value sooner.

There are, of course, many kinds of progress: scientific, technological, societal, moral, and more. And each of these has many different strands within it. A nuanced approach to advancing progress therefore involves the possibility that some of these advance more than others, which could have complex effects on the shape of the trajectory.[4]

[4] Carl Sagan (1994: 316–17) suggested that the fundamental challenge of anthropogenic existential risk stems from our technological progress outstripping our progress on becoming wiser as a civilisation. Bostrom (2014: 228–46) has a good account of the importance of advancing some kinds of progress more than others. For much more on the idea that humanity should act to advance progress (in all its nuance), one could look to the burgeoning field of Progress Studies (Collison and Cowen 2019).

But it is useful to also ask the simpler question: what if we could shift the trajectory of humanity's instantaneous value earlier by some small amount of time?[5] This may correspond roughly to advancing all forms of progress by that same amount of time. I am not meaning to suggest that all attempts to advance progress will behave like this, but that it is a clean transformation of the trajectory which captures part of what advancing progress is about, and which could be helpful in thinking about real attempts to advance progress.

We shall say that an *advancement* is a change to the default trajectory $v(\cdot)$, where the future trajectory is shifted left by some small amount of time δt. More formally:

$$v^{\star}(t) = \begin{cases} v(t), & t \leq 0 \\ v(t + \delta t), & t > 0 \end{cases}$$

(Note that this way of modelling advancements involves a discontinuous jump in the trajectory from where it was at time 0 to where it would have been at δt. I don't mean to suggest that it really would jump like this—I presume it would instead continuously rise to that level over a similar timescale to δt. But the difference this makes to the final analysis is small and not worth additional complexity.)

This definition so far leaves open the question of whether the end time also gets shifted. If τ is an exogenous end time for humanity (e.g. the sun burning out, or a collision point with a large asteroid) then we might expect it to stay fixed. But if it is endogenous to our activities (e.g. extinction via a dangerous advanced technology), we might expect it to be brought forwards by δt along with everything else. Both cases are plausible, and we'll examine them separately.

Let's begin by considering an advancement with an exogenous end time. This raises a new question of what values $v^{*}(t)$ takes between $\tau - \delta t$ and τ. I will assume that there is some standard continuation of the default trajectory $v(\cdot)$ beyond τ—what would have happened to humanity by default if we hadn't gone extinct at that time. Technically we can suppose that $v(\cdot)$ is defined and continues beyond τ, it is just that we only show (and integrate) the curve up to τ.

Given this approach, the value of an advancement is equal to the sum of the the shaded areas above the solid line in Figure 13.4 (where the new trajectory is superior) minus the sum of the shaded areas below the solid line (where the default trajectory was superior). For trajectories with many turning points, these regions can get very complex, and their magnitude unclear. But happily, this difference in values between the old and new trajectories can be approximated with a very simple expression which doesn't depend on the shape of the trajectory.

Figure 13.5 shows the shape of the default trajectory, with the dotted section at the end of the solid line showing how it would continue beyond τ for the next δt. We can divide

[5] Bostrom (2003) explores a similar approach, comparing advancements to existential risk reduction. Cotton-Barratt (2015: 9–12) models it very similarly to me, though applies the model to advancing progress in building a social movement.

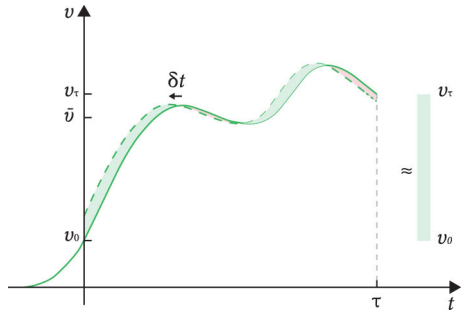


Figure 13.4 An *advancement*. The dashed trajectory represents a change to the default trajectory where each instantaneous value level is reached δt years earlier. This corresponds to shifting the default future trajectory δt units to the left. (As in all these diagrams, the scale of this shift is exaggerated here for clarity.) The difference in value between these trajectories is equal to the sum of the shaded areas between the curves, where the shaded areas above the solid line count positively and those below it negatively.

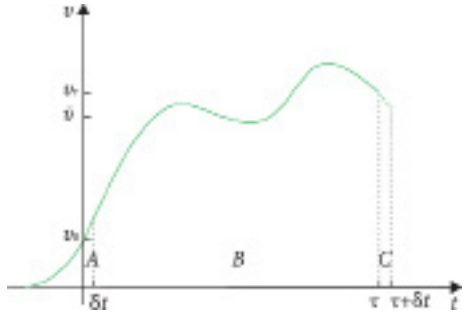


Figure 13.5 Decomposing the future trajectory into pieces that will help us find the value of an advancement.

the area under this curve into three regions, *A*, *B*, *C*. The default trajectory ends at τ. So the value of our default future, *V*, is simply given by:

$$V = A + B$$

In contrast, the alternate trajectory with the advancement skips over piece *A* and consists of pieces *B* and *C*.

$$V^* = B + C$$

So the amount by which V^* is superior to *V* is:

$$V^* - V = C - A$$

Under certain conditions, we can approximate the sizes of *C* and *A*. This depends on how quickly $v(\cdot)$ is changing at 0 and τ. If the instantaneous value is changing relatively slowly

at 0 (i.e. if $v_0 \gg |v(\delta t) - v_0|$) then we can approximate A by a narrow rectangle with height v_0 and width δt:

$$A \approx v_0 \delta t$$

Similarly for C and τ:

$$C \approx v_\tau \delta t$$

This gives a simple approximation for how much better V^* is than V (a difference we shall denote δV):

$$\delta V \equiv V^* - V \approx (v_\tau - v_0)\delta t$$

In many circumstances with a marginal advancement this will be a close approximation. For example, if the instantaneous value of humanity changes by less than 5% each year, then these approximations for A and C will be correct to within 5% for advancements of up to a year. And an advancement of an entire year would be very difficult to achieve: it may require something comparable to the entire effort of all currently existing humans working for a year. Even with a very leveraged opportunity, we might expect the kinds of advancements under consideration to be more like days than years, in which case the approximation should be accurate to better than 1 part in 1,000.

This approximation for δV has no dependence on the shape of $v(\cdot)$. It depends solely on its values at two particular times and the size of the advancement. Indeed, the value of an advancement doesn't even have any dependence on the duration of humanity's future, τ. Looking closer we can see that even the precise values of A and C depend only on the shape of $v(\cdot)$ in the immediate vicinity of 0 and τ, and don't depend on τ.

So despite an advancement being a kind of lasting change to the trajectory of humanity—a change whose effect remains at full strength over our entire future—its value doesn't scale with the length of this future. Whether humanity's future lasts a million years or a billion years doesn't affect the value of an advancement, except inasmuch as we might hope that we reach higher heights the longer we have. But if our world (or our value system) is such that the instantaneous value of humanity plateaus, then duration doesn't matter much to the value of advancements. And if our future trajectory is such that the later instantaneous values aren't much different to those today, then advancements would have very little value at all.

It is interesting to note that the kind of preference motivating advancements—for things to happen sooner rather than later—can arise even from the perfectly patient perspective we are considering in this chapter. It isn't about moving some fixed amount of value earlier in time, but about creating more value. For example, whenever the slope of the trajectory $v(t)$ is increasing over the long run, shifting this trajectory earlier in

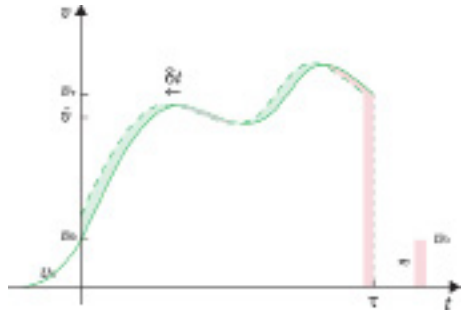


Figure 13.6 An advancement with an endogenous end time.

time makes things better on average across all subsequent times. And if the slope is also bending upwards, the amount by which things are being made better at each time is also increasing over time.

We've focused so far on the case where the end time is exogenous, but what if it is endogenous? The simplest way to model how an advancement changes an endogenous end time is to say that the end time shifts to the left by δt along with everything else (i.e. $\tau^* = \tau - \delta t$).

If so, the area under the new trajectory (V^*) will still skip over the first piece (A) in Figure 13.5, but will end before it would get the compensating benefit of the third piece (C). So the value of the future is just that of the middle piece (B) (see Figure 13.6):

$$V^* = B$$

So

$$\delta V = V^* - V = B - (A + B) = -A$$

$$\delta V \approx -v_0 \delta t$$

This is a substantial difference. Now the only real effect of the advancement is to skip the value of the next δt of humanity's trajectory. So unless humanity currently has a negative instantaneous value, an advancement with this kind of endogenous end time actually makes the future worse. In this case it really is just about moving the future benefits earlier in time rather than increasing the amount of benefits in the future.

Note that this conclusion is sensitive to the precise nature of the question we are asking. We have been comparing two different trajectories—two particular ways the world could unfold. But a natural extension of this framework could consider uncertainty by instead comparing two probability distributions over trajectories or by comparing trajectories with associated hazard curves. This could change the conclusion. For example, if we made the assumption that an advancement by δt skips over the risk in the next δt of our trajectory, this could make the value of the advancement positive again,

even when risk of ending the trajectory is endogenous (i.e. when all subsequent risk gets advanced too).[6]

In summary, when the end time is exogenous, it is easy to see how advancing progress across the board could improve our future: roughly speaking, it could replace a period at current instantaneous value with one at the final instantaneous value. This value could scale with the ultimate size of human endeavours in the future, even if it won't scale with the duration of humanity's future. But when the end time is endogenous, this kind of justification for advancing progress won't work. On the simplest model, it might just make things worse by skipping over a period at our current level of instantaneous value. So if it is to have good effects, that would need to be due to a more complex story about how it shapes the future—for example, if it skipped over some of the risk from the duration it advanced, or if it advanced progress in some areas more than others, such that it reduced existential risk or changed the trajectory in some way beyond a simple horizontal shift.[7]

Finally, note that the study of advancements also applies to their opposites: *delays*. These correspond to shifting the future trajectory to the right. They fit cleanly into the same mathematical framework by simply allowing δt to be negative.[8] Marginal delays are bad in almost exactly those circumstances when marginal advancements would be good, and vice versa.

7 Speed-ups

What if, instead of merely advancing progress, we could permanently speed it up? Perhaps there are ways of organising society that would achieve in 100 years what would have taken us 101, achieve in 1,000 years what would have taken us 1,010, and so on. As this would have a proportionally larger effect further into the future, it wouldn't be shifting the future trajectory to the left, but horizontally compressing it by some factor γ_t:

$$v^\star(t) = \begin{cases} v(t), & t \leq 0 \\ v(\gamma t(t)), & t > 0 \end{cases}$$

In this way, γ_t could be thought of as the factor by which progress is sped up, and we can call this kind of change a *speed-up*. Marginal speed-ups would correspond to γ_t being slightly greater than 1. For instance, the 1% speed-up described in the previous paragraph would correspond to $\gamma_t = 1.01$.

[6] Aschenbrenner (2020) reaches similar conclusions with his economic model.

[7] See Ord (2024) for a detailed discussion of what this means for the value of advancing progress.

[8] There is actually a small additional wrinkle, in that you need to define what $v^\star(t)$ is between $v(0)$ and $v(\delta t)$. That is, what small piece of trajectory to insert to join the past trajectory to the delayed future trajectory. The most natural answer is to have it be a flat line at $v^\star(t) = v(0)$

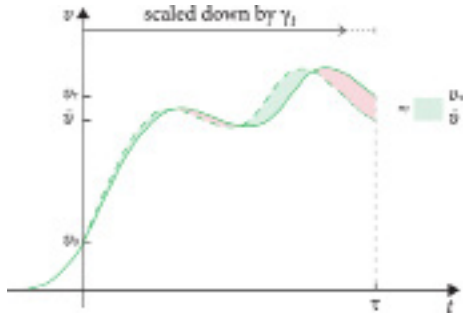


Figure 13.7 A speed-up with an exogenous end time.

As with advancements, it matters whether the end time is exogenous or endogenous. Let's first consider the exogenous case, as in Figure 13.7.

To see how this changes the value of the future, we can decompose the future into two parts. First, note that the average instantaneous value under the dashed curve is just the same as the average instantaneous value under the default trajectory: $\overline{v}$. And for marginal speed-ups (those where $\gamma_t \approx 1$), the average instantaneous value under the *dotted* line is approximately v_τ. So the new value of the future, V^*, is a weighted average of these:

$$V^* \approx ((1/\gamma_t)\overline{v} + (1-1/\gamma_t)v_\tau)\tau$$

The amount by which it is better than the default is:

$$\delta V = V^* - V \approx (v_\tau - \overline{v})(1-1/\gamma_t)\tau$$

This is a product of three factors: the amount by which the end time is better than the average time, a number just above zero representing the fraction of humanity's duration in the dotted part of the trajectory, and humanity's duration itself.

This whole effect has a positive value if and only if the instantaneous value at the end of the default trajectory is higher than its average instantaneous value.

For speed-ups with endogenous end times (as in Figure 13.8), the duration of humanity is also sped up, with $\tau^* = \tau/\gamma_t$.

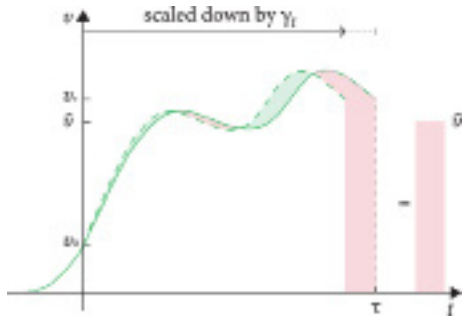


Figure 13.8 A speed-up with an endogenous end time.

The area under this new (dashed) trajectory is just V/γ_t—a compressed version of the default value of the future. And the improvement it makes on the future is:

$$\delta V = V^* - V = -\overline{v}(1 - 1/\gamma_t)\tau$$

This is of positive value if and only if the total value of the default trajectory is negative. That's because it corresponds to having all the same levels of value in the future, just spending less time at each one—making the value of the future a little smaller. So again, the distinction between whether the end is exogenous or endogenous is very important for evaluating this intervention. Speed-ups with endogenous end times aren't generically a good thing—they would need to be differentially speeding up progress or reducing risk in some way to be valuable.

How plausible are speed-ups? The broad course of human history suggests that speed-ups are possible. For example, the agricultural and industrial revolutions each seemed to substantially speed up the clock of human progress compared to what came before. And presumably they also sped it up compared to what would have happened if they had never occurred. Of course, they also had other effects, speeding up some processes more than others and causing other changes to the trajectory. But the idealised idea of a speed-up would still seem to capture a key part of what happened. Both revolutions would count as non-marginal speed-ups. More incremental speed-ups may also exist, though we wouldn't expect to be able to identify a marginal speed-up (say 1%) in the noisy historical record.

What would it look like if there were many marginal speed-ups? If there were a succession of them, happening roughly uniformly in time, the overall effect could be exponential. Imagine starting with an upwards diagonal default trajectory then at time 1 introducing a 1% speed-up, at time 2 introducing a further 1% speed-up, and so forth. After the first speed-up, the slope is 1% higher than before, after the second it is 1% higher than that (compounding). Since the slope is increasing exponentially with time, so is the curve itself, and thus so is the integral.[9]

It is plausible that some component of the exponential growth humanity has achieved across various dimensions is due to the accumulation of speed-ups like this. But it is not clear whether the key dimension—the instantaneous value of humanity—has actually grown exponentially. It probably has according to theories of population ethics where the instantaneous value is proportional to the population at that time (all things being equal), but not according to most other theories. So this case for successive small speed-ups is more suggestive than proven.

One big challenge for the idea that speed-ups are something for altruists to aim towards is that if a speed-up is possible at all, it seems likely to be overdetermined that it will happen. That is, one could calculate the difference in value between a trajectory where it doesn't happen and one where it does, leading to a high valuation—but that might not be the relevant comparison. If it is overdetermined to happen, then by making it happen now, we are just bringing forward the time when it happens. And if so, then we just have an advancement rather than a speed-up. I think this is probably the case for the agricultural revolution

[9] If each speed-up also sped up the rate at which future speed-ups would happen, the growth would be even faster, approaching a vertical asymptote at a finite time.

(which was so overdetermined it independently occurred in at least five different parts of the world (Christian 2004: 248–87)), though there is more scholarly debate about whether the industrial revolution would have ever happened had in not started in the way it did. And there are other smaller breakthroughs, such as the phonetic alphabet, that only occurred once and whose main effect may have been to speed up progress. So contingent speed-ups may be possible.

The opposite of a speed-up is a *slow-down*. It is what you get when γ_t is less than 1 and corresponds to horizontally stretching the future trajectory. Note that like speed-ups, slow-downs are defined compared to what would have happened by default. So a slow-down could take the form of an event that actively slows down future progress or it could take the form of a choice *not* to pursue some new development we were headed towards that would have sped things up. For marginal changes, slow-downs are good when speed-ups are bad, and vice versa.

8 Gains

What if, instead of adjusting the timings of the future trajectory, we could directly adjust the instantaneous value itself? An advancement is a shift of the trajectory to the left (which sometimes leads to it rising on average), but what if we could directly shift the trajectory upwards? We can all such a change a *gain* (as in Figure 13.9). Mathematically:

$$v^*(t) = \begin{cases} v(t), & t \leq 0 \\ v(t) + \delta v, & t > 0 \end{cases}$$

The value of a gain is easy to calculate. The new value is δv units higher at all times, so:

$$V^* = (\bar{v} + \delta v)\tau$$

and

$$\delta V = \delta v \tau$$

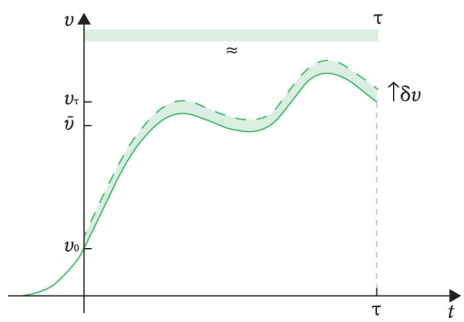


Figure 13.9 A gain of size δv.

While the idea of a gain is simple—a permanent improvement in instantaneous value of a fixed size—it is not so clear how common they are. Many kinds of permanent improvements in humanity's wellbeing might be expected to scale in value with other quality-of-life improvements or with our population. If so, they are likely to be proportional increases in humanity's value at any time, rather than fixed increases, so would not count as gains (they will be addressed next). However, certain limiting cases may count as gains. For example, a fixed improvement in everyone's quality of life in a future where the population doesn't change much would count, as would a fixed improvement in everyone on Earth's quality of life even if additional people came to exist elsewhere, or a fixed improvement in everyone's quality of life according to a system of population ethics where instantaneous value didn't scale with the size of the population at that time.

Such a fixed improvement to each individual's instantaneous wellbeing could come via a direct effect on wellbeing of fixed size, or via a proportional improvement in some instrumental good (such as income) whose effect on wellbeing is logarithmic.

Other examples could be found beyond human wellbeing. For example, a permanent improvement to the wellbeing of animals on Earth would behave like a gain (though it would require an adjustment to what $v(\cdot)$ is supposed to be representing). Or consider an improvement to a non-welfarist good, such as saving an ecosystem, a species, or a work of art. For value systems where the value of such things is proportional to how long they last, these would count as gains.

As with speed-ups, a putative gain often faces a challenge regarding whether it was truly contingent, or would have simply happened later. If the lasting benefit would have been achieved later in the default trajectory, then it is only temporary so not a true gain. For example, making a scientific discovery may make things better for all subsequent times, but if the default is (as usual) that someone else would have discovered it sometime later, then that improvement is not permanent, so not a gain. If its value lies not just in improving the value at each time, but in allowing us to get to future points in our development sooner, then it may be an advancement instead. One way that something can resist this challenge is if it takes the form of saving something irreplaceable from permanent destruction.

The opposite of a gain is a *loss*. It is what you get when δv is negative, and corresponds to shifting the future trajectory down. Losses could result from adding things of negative value, removing things of positive value, preventing things of positive value being created, or preventing the only chance that something of negative value could be remedied.

9 Enhancements

It may also be possible to permanently improve humanity's instantaneous value by a given proportion—for example, if we could make every moment across humanity's future 1% more valuable. We can call such an idealised change an *enhancement* (as in Figure 13.10), and model it as:

$$v^*(t) = \begin{cases} v(t), & t \leq 0 \\ \gamma_v v(t), & t > 0 \end{cases}$$

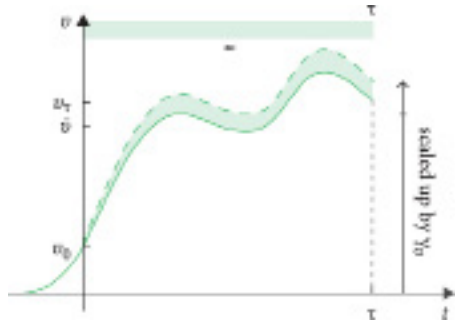


Figure 13.10 An enhancement by a factor of γ_v.

As this is simply a scaled-up version of the default trajectory, the value is trivial to calculate:

$$V^* = \gamma_v \bar{v} \tau$$

and

$$\delta V = (\gamma_v - 1)\bar{v}\tau$$

There are many kinds of changes to our future that could be modelled as enhancements—for example, improving the quality of life for everyone who will live by a modest proportion (perhaps via improvements to health, technology, prosperity, our lived environment, or our social structures). Alternatively, proportionally increasing the number of people alive at each time would be an enhancement according to certain theories of population ethics. Or improvements in our moral understanding or moral motivation could mean that we produce systematically better outcomes.

As with many of these idealised changes, they face the challenge of why this wouldn't happen eventually, even without the current effort. I think this is a serious challenge for many proposed enhancements. Those that can best resist it may be cases where there is going to be a kind of lock-in (Ord 2020: 153–8). For example, if the values that guide humanity become locked-in at some point in time, then improvements to those values prior to that point could have truly lasting impact.

The opposite of an enhancement is a *diminution*. It is what you get when $\gamma_v < 1$ and corresponds to vertically compressing the future trajectory.

10 Combinations and variations

Interventions could also produce a combination of these idealised changes. And it is easy enough to allow this in the mathematics. We simply need to keep track of all the locations where a delta or gamma could lurk and allow more than one to take a non-trivial value. The general form is:

$$v^*(t) = \gamma_v v(\gamma_t t + \delta t) + \delta v$$

The deltas can move the trajectory in any combination of up, down, left, and right. And the gammas can stretch or compress it horizontally or vertically. This allows substantial flexibility in how the trajectory is transformed. That said, the flexibility comes at a cost of increased complexity and reduced clarity about the way in which the intervention is affecting the future and why. My guess is that the individual transformations are the more useful tool.

One form of idealisation has been that these changes to the trajectory will be permanent. But of course, many changes are not. A very general way to represent them is to have a decay curve $d(t)$, that fades the effect out over time. $d(t)$ would start at 1 and monotonically decay according to some desired schedule. This could fully remove the effect by some specified time, have the effect asymptote towards zero, or have it decline to some smaller but still positive size. As an example, a temporary advancement could have the equation:

$$\upsilon^*(t) = \begin{cases} \upsilon(t), & t \leq 0 \\ \upsilon(t + d(t)\delta t), & t > 0 \end{cases}$$

Alternatively, much of the benefit of modelling temporary changes might be gained through the much simpler system of assuming the effect is at full force for some specified duration before completely vanishing. That is less realistic and less flexible but may give most of the benefit with just a single parameter.

Consider a temporary advancement where the effect ends at time t_e, after which the default trajectory resumes:[10]

$$\upsilon^*(t) = \begin{cases} \upsilon(t), & t \leq 0 \\ \upsilon(t + d(t)\delta t), & 0 < t < t_e \\ \upsilon(t + \delta t), & t_e < t \leq t_e + \delta t \\ \upsilon(t), & t_e + \delta t < t \end{cases}$$

For a temporary advancement:

$$\delta V \approx (v_e - v_0)\delta t$$

So an advancement by δt would produce the maximum benefit if it persisted until the moment of humanity's peak instantaneous value, but no further—for example, an intervention which advances progress for the entire period in which humanity is ramping up to its full and final scale in the universe, but which does nothing to change the schedule of the long denouement as the stars wind down.[11] The fact that such a temporary advancement would be superior to a permanent one may also contribute to making advancements more likely to be temporary. And while temporary advancements require an extra parameter to specify, they are simpler in some other ways: the distinction between whether the end time

[10] Advancements require this third clause to fill in part of the trajectory that would otherwise be missing.
[11] Nick Bostrom (2003) was focused on an advancement of this kind.

is exogenous or endogenous becomes moot, as does the sensitivity to the very final instantaneous value.

We could call temporary changes whose effect lasts for a time on the same scale as τ, *lasting* changes to humanity's trajectory. While any precise cut-off is arbitrary, we might operationalise this as a duration at least one tenth of the entire future of humanity.

11 Comparisons

How do these different ways of shaping the longterm future compare? (see Figure 13.11)

In the right circumstances, any of these idealised changes can be preferable to any other. It all depends on the sizes of the changes (the deltas and gammas) and the key features of the trajectory (v_0, v_τ, $\overline{v}$, and τ). But there are some important patterns that help us think about what those circumstances are and how plausible they may be.

A useful way to categorise the idealised changes is by whether their impact scales with the heights of instantaneous value we may reach (via v_τ and $\overline{v}$), the aeons we may last (τ), or both. We can see that:

- Advancements scale with $v_\tau - v_0$
- Speed-ups scale with $(v_\tau - \overline{v})\tau$
- Gains scale with τ
- Enhancements scale with $\overline{v}\tau$

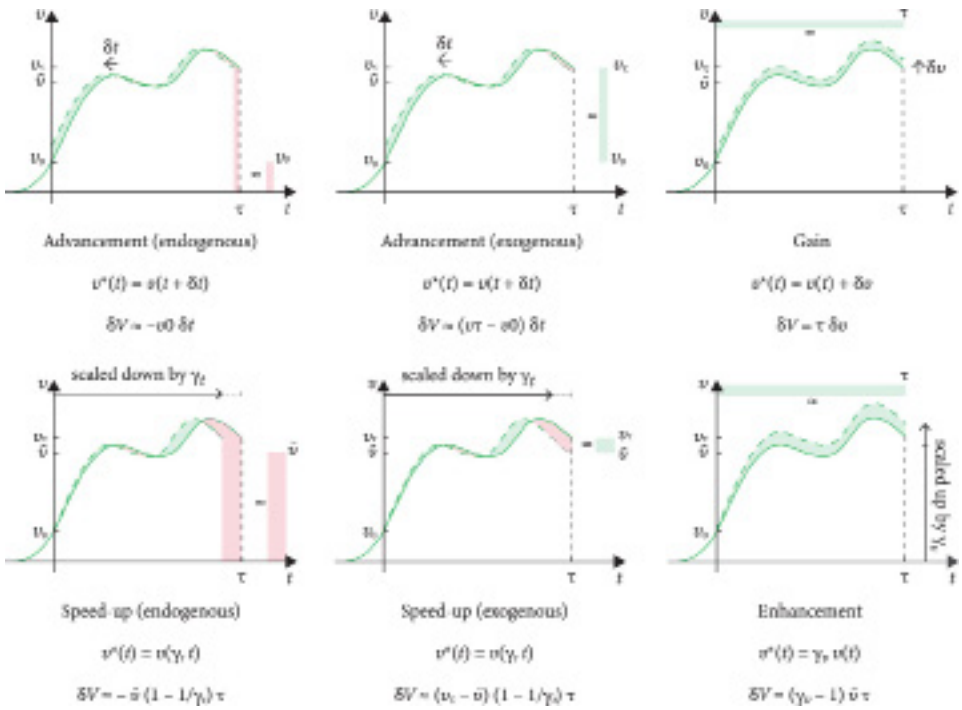


Figure 13.11 A summary of the different idealised changes to humanity's longterm trajectory.

It follows that if we consider longer and longer durations for our future, the interventions that don't scale with duration (advancements) become comparatively less important. And similarly, if we consider cases where we reach higher and higher instantaneous values, then gains become less important. Or if we start to acquire a stronger belief that the instantaneous value of the future won't be much larger than that of today, then everything but gains is lowered in importance (and advancements become especially unimportant).

Let's see how the expressions for δV can be used to better understand comparisons between these idealised changes. For example, when is an advancement better than a gain?

$$(v_\tau - v_0)\delta t > \tau \delta v$$
$$\frac{\delta t}{\tau} > \frac{\delta v}{v_\tau - v_0}$$

So an advancement is better than a gain when the proportion of humanity's lifespan that is advanced is greater than the proportion of the gap between current value and final value that is gained. While the units of years advanced and instantaneous value gained didn't directly permit comparison, we see that using this framework, the question ultimately comes down to the percentage change in each quantity, which is unitless and therefore comparable.

Moreover, the ratio by which an intervention is superior is just the ratio of these percentages. So advancing progress by a millionth of humanity's lifespan is ten times as good as permanently gaining a ten-millionth of the gap in instantaneous value between now and the end of humanity. And this perspective generalises: the best way of comparing longterm effects often comes down to a comparison of the percentage improvements.

When is an enhancement better than a gain?

$$(\gamma_v - 1)\bar{v}\tau > \tau \delta v$$
$$(\gamma_v - 1) > \frac{\delta v}{\bar{v}}$$

Here one of the biggest unknowns in the study of the longterm future—the ultimate duration of humanity's future, τ—appeared in both expressions and could simply be cancelled out. So the question of enhancements versus gains is not sensitive to this key unknown. We are again left with two percentages to compare: the percentage by which the future is being enhanced versus the percentage of the average instantaneous value that is being gained.

Some of the most important comparisons address how these idealised changes compare to reducing existential risk.

While a full accounting for existential risk unfolding over time would require us to extend the theory to deal with uncertainty (over trajectories or over τ), we can get a lot of value from a highly simplified account. We can consider that there is some probability of

existential risk occurring in our time, and model this with the risk occurring right at t=0.[12] We suppose that if this existential catastrophe were to happen, then the value of the future would be extremely small compared to V, and approximate it as zero.[13] For our present purposes of comparing longterm interventions, it doesn't matter how much risk is occurring as it shrinks the value of all the idealised interventions equally.

We can then consider an intervention that increases our probability of surviving this near-term existential risk by a factor, γ_p. For instance, if there were 20 percentage points of near-term existential risk (so an 80% chance of survival), and the intervention increased that survival chance by 1 percentage point, then $\gamma_p = 81/80 = 1.0125$. On this model:

$$\delta V = (\gamma_p - 1)\bar{v}\tau$$

Reducing near-term existential risk is thus another kind of intervention that scales with both $\bar{v}$ and τ. Indeed, the effect it has on the (expected) value of the future is almost identical to that of an enhancement: they both multiply the entire value of the future by some factor. So if prioritising between an enhancement and existential risk reduction, it all comes down to which one has the higher factor.

Though remember that for something to be a genuine enhancement, it needs to be truly contingent—a kind of permanent proportional improvement that would never have happened otherwise. Part of the reason reducing existential risk is so important is that it is much easier to meet this bar for contingency—that the loss is irreversible (or nearly so) is built into the definition of existential risk.

We can also compare existential risk to other idealised changes. When would an advancement (with an exogenous end time) be better than lowering existential risk?

$$(v_\tau - v_0)\delta t > (\gamma_p - 1)\bar{v}\tau$$

$$\frac{\delta t}{\tau} > (\gamma_p - 1)\cdot\frac{\bar{v}}{v_\tau - v_0}$$

This is a bit harder to interpret, but not impossible. In some situations, we might expect $\bar{v}$ and $(v_\tau - v_0)$ to be of similar magnitude (e.g. if humanity's instantaneous value spends a long time at a high plateau). In that case, advancements are better roughly when $\delta t/\tau > (\gamma_p - 1)$—when the percentage of the duration of humanity's future that we advance is greater than the percentage by which our survival probability is increased. This makes it difficult for advancements to beat existential risk reduction in scenarios where humanity would have a long lifespan if only it could survive the near-term risks. For example, on a million-year lifespan (that of a typical species) a one-year advancement would be roughly as important

[12] So long as the timeframe for 'our time' is less than one percent of τ, this way of modelling near-term risk will be fairly accurate.

[13] This is less problematic than it sounds. If the remaining value were, say, a tenth that of V (about the highest remaining value that could still count as an existential catastrophe), then the approximation as zero value still wouldn't change the results much. It would just mean that the true δV for existential risk reduction was 10% smaller than the approximation.

as a one-in-a-million improvement in nearterm survival probability—but the latter seems much more achievable.

Even if the simplifying assumption that $\overline{v}$ is similar in scale to $(v_\tau - v_0)$ were not true, it is still very difficult for advancements to beat existential risk reduction, as in order to compensate, this ratio would need to be extreme. It would require a trajectory where the final value was far higher than the average, or a temporary advance up to an intermediate time at the top of a very high and narrow peak in instantaneous value.

It is important to remember that all these equations and comparisons are just for the pure, idealised, changes. Real attempts to improve progress would undoubtedly shift some areas more than others, leading to more complicated effects on the shape of our future. But one upshot of this analysis is to find the situations in which these more complex effects would be necessary.

It seems to me that for attempts to advance progress to be more valuable over the long term than attempts to reduce existential risk, such complex effects would be required. For example, advancing defensive technologies or collective wisdom may reduce existential risk, and advancing moral progress may lead to better values eventually becoming locked in and guiding the future. Even if these kinds of effects might initially seem to be second-order, the fact that they can also scale with τ suggests that they could dominate the longterm value of attempts to advance progress.

12 Conclusions

In this essay, we have explored a quantitative framework for modelling the longterm trajectory of humanity. By tracking the instantaneous value over time, we've enabled quantitative evaluations and comparisons of different trajectories in terms of the area under the curves. The analysis revealed four different kinds of idealised change to the trajectory corresponding to vertical and horizontal shifts and stretches. Even relatively small changes of these kinds might have vast effects on the value humanity achieves over all time. And given some approximations, these can be usefully analysed and compared. In particular, we've seen that it can be difficult for a pure advancement or gain to rival those interventions like speed-ups, enhancements, and existential risk reduction that scale with both the instantaneous value of the longterm future and its duration. And we've seen that the value of both advancements and speed-ups critically depend on whether they shift the end time as well.

The primary aim has been to help develop a theoretical underpinning for understanding longtermist interventions, but I hope that it will also be of some practical use in finding and comparing tractable interventions of these kinds. And it may also help us weigh the longterm effects of everyday actions taken by people and governments. While such actions are usually aimed at short-term effects, given the amounts of value at stake over the long term, it is possible that their ultimate impacts are often driven by their longterm effects. If so, this framework could help us better understand their impacts.

All of this was done without discounting. This was made possible by parameterising the lifespan of humanity (which can be treated as exogenous or endogenous) and showing how the value of different interventions scales as a function of this parameter. It also involves explicit modelling of the longterm impacts of our actions, since (unlike in traditional economic analyses) these impacts no longer vanish. While the framework is fully general in

terms of the shape of the trajectory, we saw that the values of these idealised changes were remarkably independent of the details of this shape and depended only on a small number of key parameters, greatly helping simplify the analysis.

The framework has substantial room for further development. An important extension is to add uncertainty, either by the fully general approach of comparing probability distributions over trajectories, or by associating a hazard curve with each trajectory. Another extension would be to consider other kinds of changes to the trajectory, including non-marginal changes.

And one could also consider a set of idealised shapes that the future trajectory might take. For instance, what more could we say if we knew the shape was a rapid rise to a long plateau? Or what if there was an exponential rise in instantaneous value all the way to the end time? Such an exponential trajectory has the special feature that an advancement would act just like an enhancement—making every future moment proportionally better. However, despite the recent centuries of exponential *economic* growth, there is reason to be doubtful of exponential growth of intrinsic value, especially over such very long timescales (Ng 1991). Consideration of physical limits to growth might instead suggest an idealised trajectory that grows as a cubic—representing the longterm growth in humanity's resources were we to spread out through the universe.[14] Collating a set of such idealised shapes for trajectories and exploring how they differ may help us better understand how longtermist priorities depend on the broad shape of the future.[15]

References

Adams, F. C., and Laughlin, G. (1997), 'A Dying Universe: The Long-Term Fate and Evolution of Astrophysical Objects', in *Reviews of Modern Physics* 69/2: 337–72.

Adams, F. and Laughlin, G. (1999), *The Five Ages of the Universe* (Free Press).

Aschenbrenner, L. (2020), 'Existential Risk and Growth', GPI Working Paper No. 6-2020 (Global Priorities Institute, Oxford University).

Barnosky, A. D. et al. (2011), 'Has the Earth's Sixth Mass Extinction Already Arrived?', in *Nature* 471/7336: 51–7.

Baum, S. D. et al. (2019), 'Long-Term Trajectories of Human Civilisation', in *Foresight* 21/1: 53–83.

Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future*, PhD thesis, Rutgers University: especially 69–73.

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', in *Utilitas* 15/3: 308–14.

Bostrom, N. (2011), 'Infinite Ethics', in *Analysis and Metaphysics* 10: 9–59.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4/1: 15–31.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

Broome, J. (2004), *Weighing Lives* (Oxford University Press).

[14] Note that the choice of cubic growth makes a tacit assumption that each new location provides its own stream of value (perhaps as a location for our descendants to live, or due to the energy flow of starlight at that location). But it is also possible that the contribution of new locations is better thought of as a fixed set of resources (e.g. the materials or energy at that location). If so, it might be better to think of the longterm growth as quadratic. It has also been suggested (Sandberg et al. 2016; Ord 2021) that the ultimate physical limits may be set by a civilisation that expands to secure resources but doesn't use them to create value until much later on, when the energy can be used more efficiently. If so, one could tweak the framework to model this not as a flow of intrinsic value over time, but a flow of new resources which can eventually be used to create value.

[15] I'd like to thank Nick Beckstead for early discussions about quantifying the value of advancements, and Owen Cotton-Barratt, Will MacAskill, and Finlay Moorhouse for many further conversations that helped this framework come together. This work was funded by a grant from Open Philanthropy.

Cassan, A. et al. (2012), 'One or More Bound Planets per Milky Way Star from Microlensing Observations', in *Nature* 481/7380: 167–9.
Christian, D. (2004), *Maps of Time: An Introduction to Big History* (University of California Press).
Collison, P. and Cowen, T. (2019), 'We Need a New Science of Progress', *The Atlantic*, 30 July.
Cotton-Barratt, O. (2015), 'How Valuable Is Movement Growth?', working paper (Centre for Effective Altruism).
Frederick, S. (2003), 'Measuring Intergenerational Time Preference: Are Future Lives Valued Less?', in *Journal of Risk and Uncertainty* 26/1: 39–53.
Galway-Witham, J. and Stringer, C. (2018), 'How Did *Homo sapiens* Evolve?', in *Science* 360/6395: 1296–8.
Greaves, H. and MacAskill W. (2021), 'The Case for Strong Longtermism', GPI Working Paper No. 5-2021 (Global Priorities Institute, Oxford University).
Harrod, R. F. (1948), *Towards a Dynamic Economics: Some Recent Developments of Economic Theory and Their Application to Policy* (Macmillan and Co.).
Koopmans, T. C. (1960), 'Stationary Ordinary Utility and Impatience', in *Econometrica*, 28/2: 287–309.
MacAskill, W. (2022), *What We Owe the Future* (Basic Books).
Ng, Y.-K. (1991), 'Should We Be Very Cautious or Extremely Cautious on Measures That May Involve Our Destruction?', in *Social Choice and Welfare* 8: 79–88.
Ng, Y.-K. (2005), 'Intergenerational Impartiality: Replacing Discounting by Probability Weighting', in *Journal of Agricultural and Environmental Ethics* 18: 237–57.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Ord, T. (2021), 'The Edges of Our Universe', *arXiv:2104.01191*.
Ord, T. (2024), 'On the Value of Advancing Progress', https://www.tobyord.com/writing/progress
Pigou, A. C. (1920), *The Economics of Welfare* (1st ed.) (Macmillan and Co.).
Pivato, M. (2008), 'Sustainable Preferences via Nondiscounted, Hyperreal Intergenerational Welfare Functions', MPRA Paper No. 7461.
Ramsey, F. P. (1928), 'A Mathematical Theory of Saving', in *The Economic Journal* 38/152: 543.
Sagan, C. (1994), *Pale Blue Dot: A Vision of the Human Future in Space* (Random House).
Sandberg, A., Armstrong, S., and Cirkovic, M. (2016), 'That Is Not Dead Which Eternal Lie: The Aestivation Hypothesis for Resolving Fermi's Paradox', in *Journal of the British Interplanetary Society* 69: 406–15.
Stern, N. H. (2007), *The Economics of Climate Change: The Stern Review* (Cambridge University Press).

14
Longtermism and Cultural Evolution

Aron Vallinder

1 Introduction

Humans are rather unique animals. We occupy a wider range of habitats than any other large land animal (Boyd and Richerson 2009). We have modified more than a third of the Earth's land area (Vitousek et al. 1997). We have used science and technology to radically transform and increasingly understand our environment. We write poetry and compose symphonies. We have religions, social norms, international agreements, courts of law, and a wide range of other institutions. What makes us unique is in large part our capacity for cumulative cultural learning. That is, by observing others we are able to learn skills, norms, beliefs, and behaviours that have taken shape over several generations. This means that individuals don't have to reinvent the wheel, but can instead rely on insights and innovations that have accumulated over time. Indeed, most of the things we take for granted in contemporary society would be impossible without cumulative culture: no single individual could discover them on their own.

According to the field of cultural evolution, cultural traits are subject to Darwinian evolution just like genes are, and can be studied using similar types of models. This process has been a major driver of human history so far. Unless something unprecedented happens, we should expect it to continue to shape our trajectory into the long-term future. For this reason, the field of cultural evolution provides a set of tools and insights that can inform our thinking about the long-term future. For example, it can help us identify various possible cultural selection pressures an intervention aimed at influencing the long-term future must be able to survive in order to persist over time. The field of cultural evolution should therefore be of significant interest to longtermists, i.e. those who (as I will here understand the term) believe that our impact on the long-term future should be a major moral consideration today. Moreover, some have proposed cultural change—and in particular values change—as a potentially promising way of having significant impact on the long-term future (Anthis and Paez 2021; MacAskill 2022). The tools of cultural evolution can help us evaluate the feasibility and desirability of such interventions. In the next section, I provide an introduction to cultural evolution: how it operates on an individual level, the importance of cumulative cultural evolution and how it came about, and how intergroup competition has shaped our history, including in the emergence of large-scale cooperation. After this, in section 3, I explore how cultural evolutionary considerations can inform our thinking about the long-term future. I suggest that such considerations provide reason against 'locking in' certain values or other features of society indefinitely, or that at the very least should make us wary of doing so prematurely. I argue that cultural evolution supports an increase in experimentation and variation. I also examine how cultural selection pressures may constrain the range of feasible scenarios for the long-term future. Overall,

Aron Vallinder, *Longtermism and Cultural Evolution* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Aron Vallinder 2025. DOI: 10.1093/9780191979972.003.0014

I hope that these explorations will demonstrate that cultural evolution is an underappreciated source of insights relevant to the project of trying to steer the course of the far future and that further research in this direction would be useful.

2 The basics of cultural evolution

'Culture' in the relevant sense is a broad category: any information that is transmitted via social learning mechanisms such as teaching and imitation. This includes languages, hunting practices, cooking techniques, programming languages, religious beliefs and rituals, as well as social norms, values, preferences, and a range of other things. How do cultural traits arise, spread, and change over time? For evolutionary adaptation to happen, three conditions must be met:

1. *Variation*. Entities vary in their characteristics.
2. *Inheritance*. Characteristics that contribute to differential fitness are heritable (i.e. there is transmission of information).
3. *Differential fitness*. Entities with different characteristics have different rates of survival and reproduction.

Genes and cultural traits are two types of entities that meet these conditions. In the case of culture, consider for example the length of arrows used by hunters in some community. Likely there will be variation in this cultural trait: not all hunters use the exact same length. Second, some arrow lengths will lead to greater hunting success than others, increasing survival rates for hunters using them. In turn, this success makes other individuals more inclined to copy them, further increasing their spread. Or consider norms around food sharing. Communities with more cooperative and generous food-sharing practices may experience increased social cohesion and support, and ultimately higher survival and reproduction rates. As a result, these food-sharing norms may become more prevalent.

2.1 The mechanisms of cultural micro-evolution

As the examples of arrow length and food sharing norms illustrate, cultural evolution differs from genetic evolution in some key ways. For example, genes are almost always inherited from one's biological parents (vertical transmission), whereas cultural traits can also be acquired from a range of other sources, such as other members of the parental generation (oblique transmission), other members of one's own generation (horizontal transmission), or members of later generations. Horizontal transmission allows for adaptive cultural traits to spread faster than does vertical transmission (Cavalli-Sforza and Feldman 1981: 351–357). In small-scale societies, oblique and horizontal transmission typically happens on a one-to-one basis, but industrial society allows for one-to-many transmission via mass media, which also has the effect of speeding up cultural evolution.

Moreover, genes are discrete units of information whereas many cultural traits (such as arrow length) are continuous. Genetic information is copied and replicated, but cultural evolution does not always require a notion of replication. For example, suppose that an

individual decides on arrow length by averaging the arrow lengths of the three most successful hunters. There's no obvious sense in which anything was replicated, but it's nevertheless clear that something was inherited (Boyd 2002).

Within this framework, several factors affect how cultural traits evolve over time. First, consider the sources of variation. New traits can be introduced through either random variation or guided variation. In random variation, cultural traits mutate just like genes. These variations can arise from mistakes in learning or the transmission of information, or from spontaneous innovations. In guided variation, by contrast, individuals actively modify, adapt, or innovate cultural traits based on their own experience, knowledge, or reasoning before passing them on to others.

Second, consider how the distribution of existing cultural traits changes over time. Even without cultural learning, cultural traits that increase the fitness of their bearers will increase in frequency as a result of natural selection. The distribution of cultural traits can also change over time as a result of cultural drift, i.e. random fluctuations in the frequency of cultural traits due to sampling errors in the transmission process. In small populations, cultural drift can have a significant impact on the distribution of cultural traits, leading to the loss of some traits and the fixation of others over time. But perhaps the most important force behind changes in frequency over time is cultural selection, or biased transmission. Cultural learners don't learn at random, but rather show preferential adoption of certain cultural traits based on factors such as frequency, model, or content. Let's consider these factors in turn.

In the most common type of frequency bias, conformist transmission, we tend to preferentially copy the most common trait. Suppose that you want to buy a new pair of headphones. To help you decide, you check out five different product recommendation websites and find that three of them recommend the same pair of headphones. If you would pick the product recommended by three out five websites more than 60% of the time, you are engaging in conformist transmission. Models of conformist transmission have shown it to be adaptive in a wide range of circumstances. In a model by Richerson and Boyd (1985), the environment varies spatially, so that which cultural trait is adaptive depends on one's location. In each environment, however, learning mechanisms cause the adaptive cultural trait to be more widespread than the other. At the same time, migration introduces less favoured variants. In this setting, individuals predisposed to acquire the most common trait will be more likely to acquire the favoured trait. Henrich and Boyd (1998) extend this result, showing that conformist transmission is also adaptive in environments that are both temporally and spatially variable. Nakahashi, Wakano, and Henrich (2012) point out that many models of conformist transmission focus on the case where there are only two cultural variants, and argue that conformist transmission only becomes more adaptive the more traits there are. Conformist transmission has also been experimentally well documented (Muthukrishna, Morgan, and Henrich, 2016). Henrich and Boyd (1998) show that conformist transmission supports the emergence of stable group differences that persist over time.

In model-based bias, learners preferentially adopt traits from certain kinds of individuals. If the skill in question is easy to assess, you can simply try to emulate the most successful individual. This success bias is well established, and has been observed both in laboratory settings and in the real world (Henrich and Gil-White 2001; Henrich and Henrich 2007: ch. 2; Henrich and Broesch 2011). However, in many cases there may not be an unambiguous measure of success or any other way to easily discern the most highly skilled individual.

In those cases, one useful strategy can be to observe whom others pay attention to, defer to, and imitate. Since other people have faced the same challenge of figuring out who to learn from, you can take advantage of the efforts they have already made. Prestige bias refers to this tendency to preferentially learn from and imitate others who are perceived as having high status, success, or prestige within a social group. This bias guides individuals to acquire cultural knowledge, skills, and behaviours from those who are considered to be the most knowledgeable, experienced, or influential. Prestigious individuals benefit from increased social standing and influence, and those who defer to them benefit from preferential access affording them greater learning opportunities. While prestige bias is often adaptive, we also have a tendency to overimitate, such as when deciding what product to buy based on celebrity endorsement.

In content-biased transmission, there is a preference for acquiring specific types of information or cultural traits based on their inherent characteristics. Only some cultural information enhances fitness. Natural selection has therefore favoured paying greater attention to certain kinds of cultural information, such as information about kinship, social norms, reputation, and animals and plants (Chudek, Muthukrishna, and Henrich 2015).

2.2 Cumulative cultural evolution

Boyd and Richerson (2005: ch. 1) construct a model to explore when social learning (i.e. learning from other individuals rather than on one's own) is adaptive. Suppose some population lives in an environment that can be either wet or dry. If the environment is dry, hunting and gathering is the best strategy. If on the other hand it is wet, farming is the best strategy. In deciding which strategy to pursue, individuals have two sources of information to consider: individual learning (i.e. their own observations) and social learning (i.e. copying a member of the previous generation). Neither source of information is perfect. Individual learning leads to the right answer on average, but sometimes leads astray. For example, even a dry environment might see a series of consecutive rainy years, leading those relying only on individual learning to mistakenly adopt farming. Social learning can correct for these errors in individual learning. However, if you have moved to a new environment, imitating the previous generation may lead you astray. If individuals do not frequently move between different environment types, the best learning strategy is to mostly imitate, only relying on individual learning when it is highly accurate. On the other hand, if individuals move so frequently that their environment is effectively random with respect to that of the previous generation, social learning adds no value, and the best strategy is to rely on individual learning alone. In between these extremes, some mix of individual and social learning is best.

Many other animals are capable of social learning. For example, chimpanzees have been observed using tools, such as sticks, to extract termites from their nests. They learn this behaviour by watching and imitating other chimpanzees (Boesch and Boesch 1990). Bottlenose dolphins have been documented using marine sponges to protect their noses while foraging on the ocean floor. This behaviour is learned from their mothers and passed down through generations (Mann et al. 2012).

But a crucial feature of human cultural evolution is that it is cumulative. Cumulative cultural evolution is when culturally transmitted traits become so complex that no single

individual could design them on their own, via asocial learning alone (Boyd and Richerson 1996). While there is evidence of some cumulative cultural evolution in non-human animals like chimpanzees (Yamamoto, Humle, Tanaka 2013) and New Caledonian crows (Hunt and Gray 2003), no other species relies on a complex body of accumulated cultural information to survive to the extent that humans do. Henrich and Muthukrishna (n.d.) argue that this cumulative cultural evolution is what explains humanity's dominance. Our position is not due to the intelligence of the individual, but rather to the cultural knowledge that has accumulated over generations.

The cumulative nature of cultural evolution is vividly illustrated by the many stories of lost European explorers reported by Henrich (2015). For example, in 1860 a small group of explorers travelling across the interior of Australia ran out of provisions and were forced to live off the land. Eventually, they made contact with a local Aboriginal group who shared food with them, including bread made from the nardoo plant. After this encounter, the explorers managed to find the plant themselves, pound the seeds, make flour, and bake nardoo bread. Initially it seemed like they had come across a reliable source of calories, but progressively they became weaker, with some of them dying from starvation. It turned out that nardoo is indigestible and mildly toxic unless properly processed. In order to make it edible, the Aboriginal individuals followed an elaborate procedure of preparation. This procedure had taken shape through trial and error over many generations. The European explorers, by contrast, did not have access to this culturally evolved knowledge, and instead faced what turned out to be the insurmountable task of figuring it out for themselves.

The power of cumulative cultural evolution has also been explored in laboratory studies. Muthukrishna et al. (2014) asked participants to carry out a difficult task that they had no previous experience with. Participants were arranged into 10 generations of five people each, with information sharing between generations. In one treatment, participants in generations 2 to 10 had access to guidance from all participants in the previous generation. In the other treatment, participants only had access to information from one of the five participants of the previous generation. The treatment with more cumulative cultural evolution led to significantly greater performance.

If cumulative cultural learning provides such an adaptive advantage, why is it not more widespread? Why did it only emerge in the past couple of million years? Various researchers have proposed that for cumulative cultural evolution to get off the ground, certain cognitive, behavioural, or other preconditions are necessary. For example, Tomasello, Kruger and Ratner (1993) argue that high-fidelity transmission is necessary for traits to persist over time, and that this high accuracy is cognitively demanding, and therefore is only possible in animals of sufficient cognitive development. While some dispute the importance of high-fidelity transmission for cumulative cultural learning (e.g. Sterelny 2021: 9), most researchers still agree that complex cognition is nevertheless a necessary precondition for one reason or another. Consistent with this hypothesis, average encephalization (brain size relative to body size) in mammals has increased over the past 66 million years (the Cenozoic era). In the human lineage, the expansion has been particularly fast over the past few million years. Five million years ago, our ancestors had a brain volume of around 350cm^3, compared to the 1,350cm^3 of modern humans. Most of this increase (from 500cm^3) happened in the past 2 million years. What accounts for this fast development? Boyd and Richerson (2005: ch. 4) suggest that changing climatic conditions played a crucial role. Over the past several million years, the climate became much more variable than previously, with the effect of making

existing habitats change and become less stable. This led to increased selection for abilities to cope with more variable environments, which includes more complex cognition.

Heyes (2018) agrees that advanced cognitive capacity is a precondition for cumulative culture, and suggests two additional factors. First, relative to other primates, humans are remarkably peaceful and tolerant of others, including strangers. This creates an environment where individuals are exposed to more potential models to learn from. Second, we are endowed with various attentional biases that guide cultural learning. Almost from birth, these biases guide us to look at human faces and listen to human voices, thereby facilitating teaching and learning.

Whatever the exact set of preconditions are, once cumulative cultural evolution got started, it is likely that cultural learning harnessed various other of our psychological traits, and that these traits were in turn shaped to facilitate greater and more efficient learning. For example, Tomasello (2000) and Boyd and Richerson (2005) suggest that theory of mind, i.e. our ability to infer the beliefs, desires, and intentions of others, proved useful for learning skills and behaviours from others. If I know what someone is trying to do, I can more easily copy them. Theory of mind may initially have evolved because it allowed people to better predict the behaviour of others in their social group. Once it emerged, it was able to support observational learning and cumulative cultural evolution.

2.3 Large-scale cooperation and intergroup competition

Cumulative cultural evolution likely emerged sometime in the last few million years. For most of this time, innovations were slow to spread. The earliest known stone tools date from 3.3 million years ago. So called Oldowan stone tools date from around 2.5 million years ago, and the more advanced Acheulian tools are from 1.8 million years ago. It is only in the past few hundred thousand years that innovations take less than 100,000 years to become established (Sterelny 2021). In particular, starting around 12,000 years ago, the pace of cultural evolution picked up dramatically. As agriculture emerged, humans began cooperating in larger and more hierarchical groups. Over time, there has been a dramatic increase in the scale of cooperation among humans. From nuclear families to nation states and beyond, how did this happen?

There are some genetic mechanisms that foster small-scale cooperation, such as kin-based altruism ('help your relatives') and direct reciprocity ('if you help me I help you'). However, to scale up cooperation further, other mechanisms are needed. Evolutionary theorists have studied how mechanisms involving reputation, punishment, and signalling can support the emergence of large-scale cooperation. For example, cooperation can be sustained in models of diffuse punishment, where those who defect can be punished (at some cost) by any punishers in the group. However, this creates a second-order free-rider problem: who will punish punishers that refrain from punishing to evade the cost? Diffuse punishment can also serve as a way for punishers to signal their prosociality (cooperativeness and trustworthiness), thereby increasing their own chances of favourable social interactions in the future. However, it turns out that these mechanisms can support any equally costly action even if it doesn't benefit anyone (Boyd and Richerson 2009: 3283). Therefore, while these mechanisms can explain how cooperation is sustained over time, they cannot explain how that behaviour (and not some other equilibrium) arises in the first place.

Boyd and Richerson (2009) argue that intergroup competition played a crucial role in the emergence of large-scale cooperation. To understand how intergroup competition works, we can think of individuals as belonging to nested hierarchies of social groups. For example, they might belong to nuclear families, which are united into clans, which are in turn united into tribes. Nuclear families that unite into clans tend to outcompete independent nuclear families, for example by being at an advantage should any violent conflict arise. Groups with social norms that are more conducive to large-scale cooperation will be more successful, and such norms will therefore spread. Sometimes the interests of lower-level groups may not be aligned with those of the larger unit, such as when clans compete for power and influence within a tribe. Greater cooperation at lower levels (e.g. nepotism) can be deleterious for cooperation at higher levels. As a result, groups that better manage to suppress damaging low-level cooperation may enjoy a competitive advantage.

Henrich (2015) identifies five important mechanisms of intergroup competition: violent conflict, varying group survival rates, migration, fertility rates, and prestige-biased group transmission.

1. *Violent conflict*. Violent conflict is perhaps the most vivid form of intergroup competition. War, raiding, and other violent conflict can result in the elimination or assimilation of weaker social groups by others who have norms and institutions that are more conducive to cooperation, or have other competitive advantages. Some have argued that warfare facilitated transitions to larger scales of cooperation and social complexity (Choi and Bowles 2007; Morris 2014; Turchin 2016).
2. *Varying group survival rates*. In hostile environments, only groups with a sufficient level of cooperation and sharing will be able to survive and grow, and those norms that promote such behaviour tend to become more prevalent. For example, Stark (1996) argues that norms of care gave Christians higher survival rates than the rest of the Roman population during both the Antonine plague (AD 165–180) and the plague of Cyprian (AD 249–262), thereby contributing to Christianity's rise from AD 40 (1,000 Christians) to AD 350 (34 million Christians), growing at about 40% per decade for three centuries.
3. *Migration*. Given that social norms can create groups with higher well-being and quality of life, many people will want to emigrate from less successful groups to more successful ones. Immigration can also serve to increase cultural variation, potentially spurring greater innovation.
4. *Fertility rates*. Social norms, such as religiously prescribed pro-fertility norms, can affect a group's fertility rate. Given that children will typically come to share their group's norms, cultures with pro-fertility norms will become more widespread over time. Stark (1996) argues that another important factor in Christianity's rise was its ban on female infanticide, a widespread practice in the Greco-Roman world. Together with generally higher birth rates, this made the Christian population grow at a substantially faster pace than the rest of the Roman world. More recently, Kaufmann (2010) has suggested that although birth rates are generally declining across the world, some religious groups appear to be resisting the trend. More specifically, he argues that groups like Mormons, the Amish, Hutterites, Salafist Muslims, and Haredi Jews have not only very high birth rates, but also sufficiently high rates of retention

that they are growing at substantially faster rates than other groups. However, a subsequent assessment found that growth rates may be declining for some of these groups (Juniewicz 2022).

5. *Prestige-biased group transmission*. People tend to pay greater attention to individuals from more successful groups, e.g. groups with higher living standards. For example, new nations may take inspiration from the constitutions or broader set of institutions of successful countries.

Intergroup competition has plausibly played a major role in shaping the course of human history. For example, Scheidel (2019) argues that one crucial reason why the Industrial Revolution happened in Europe and not in e.g. China was the fact that since the fall of the Roman Empire, Europe—unlike most other parts of the world—was never united into a single empire, instead consisting of several smaller units of roughly equal power. This made intergroup competition a much more important selection pressure, driving further development and innovation. Mokyr (2016) also emphasizes the fragmented nature of Europe in his account of how a cultural shift in the early modern period facilitated the Industrial Revolution. Fragmentation meant that European rulers were competing with one another for the most skilled citizens, be they painters, artisans, musicians, or engineers. Such competition between states also made it more difficult for those defending conservatism to coordinate their attempts to suppress intellectual innovators. Those who were persecuted in one state could often set up shop in another one instead. This was in part made possible by Europe's unusual combination of political fragmentation with intellectual and cultural unity—an integrated market for ideas. This unity came from Europe's classical heritage, the use of Latin as lingua franca, and the Christian Church. It allowed for the emergence of The Republic of Letters, a transnational community of scholars who disseminated ideas and corresponded with one another, giving intellectual innovators a much larger audience than they could otherwise have had. It also provided a set of institutional incentives that encouraged academic superstars and allowed heterodox scholars to spread their original ideas in the hope of gaining prestige. Among these scholars the idea that it was both possible and desirable to understand, manipulate, and improve upon the natural world began to take hold. Mokyr argues that this 'culture of growth' played a crucial role in enabling the Industrial Revolution.

3 Longtermist lessons from cultural evolution

How can the study of cultural evolution help to guide the project of trying to steer the course of the far future? At the most general level, for as long as competition between cultural units (nations, religions, ideologies, firms, subcultures, etc.) remains a potent force in shaping the trajectory of Earth-originating life, the tools of cultural evolution can help us gain a better understanding of what the long-term future may look like and what, if anything, we can do to influence it. For some particular change to persist over long time spans, it must be able to successfully compete and survive the process of cultural evolution over that time span. In this way, cultural evolution can help us assess the feasibility of various proposed longtermist interventions.

Consider for example the suggestion to evaluate longtermist interventions based on the significance, persistence, and contingency of the states of affairs those interventions are likely to bring about (MacAskill 2022; MacAskill, Thomas, and Vallinder 2022). In this framework, the significance of a state of affairs is its average value per unit time. To calculate its total value, we also need to know its persistence, i.e. how long it lasts for. When evaluating longtermist interventions we also care about contingency, i.e. to what extent the state of affairs can be traced back to some particular decision or other originating event. If an intervention brings about change which, though highly persistent, would have happened soon after even without the intervention, its longtermist value is correspondingly smaller. Cultural evolution can inform our thinking about these factors, particularly persistence and contingency. To persist, a trait must be able to survive the process of competition. To be contingent, it must be the case that competitive pressures would not have brought it about sooner or later anyway.

With this framework in mind, many proposed longtermist interventions fall into two broad categories. One set of interventions aims to reduce the risk of human extinction, whether by unaligned artificial intelligence (AI), engineered pandemics, or some other global catastrophe. Given some assumptions (e.g. that survival is a net positive, and that the expected lifespan of humanity conditional on this risk reduction is not very short), it's clear how such interventions may score highly across all three dimensions. Assuming that the risk reduction happens in the near term, we don't need to consider the dynamics of cultural evolution in order to explain how persistent influence is possible.[1]

Another set of interventions aims to increase the value of the long-term future conditional on a long future containing a large number of sentient and intelligent beings. In this case, the path to long-term impact is less clear than it is for extinction risk mitigation. One may reasonably worry that the effects of any such intervention will eventually wash out, with no impact on the long-term future. Moreover, even if it does have some long-term impact, how sure can we be that it is in fact for the better? In response, Greaves and MacAskill (2023) suggest that there may be certain *persistent states* such that once the world enters a persistent state, it will remain in that state for a very long period of time (in expectation at least). If we can influence which persistent state the world enters, we can thereby have predictable impact on the long-term future. Human extinction is one clear example of a persistent state: if humanity went extinct, it is plausibly unlikely that another species that could realize humanity's potential would evolve. But there may be other persistent states as well.

3.1 Artificial general intelligence and lock-in

One salient possibility is that artificial general intelligence (AGI) that greatly exceeds human performance in most areas of interest could allow for indefinite value lock-in. Finnveden, Riedel, and Shulman (2022) argue that AGI will enable precise preservation of goals into the far future, the creation of institutions that intelligently pursue those goals, and

[1] At least so far as the first-order effects are concerned. In theory it could be the case that interventions that reduce risks today have the further effect of making future generations less inclined to reduce those risks, but it's unclear whether we have any reason to think that this is ever true in practice.

the prevention of any disruption to its pursuit, be it natural catastrophes or other agents. Let's consider these in turn.

1. *Preserving information*. Digital error correction can ensure that information describing the goals can persist into the long-term future, and storing this information redundantly in several places further increases persistence.
2. *Executing intentions*. Ensuring that the goals are pursued as intended requires solving the AI alignment problem. Many have claimed that this is an exceptionally difficult problem (e.g. Bostrom 2014; Ngo 2020; Cotra 2021), but assuming it can be solved, we would presumably end up with a system very well equipped to execute programmed intentions over the long-term future. Moreover, if we fail to solve the AI alignment problem, one plausible scenario is that we still end up with value lock-in, only that now the locked-in values are those of an unaligned AI rather than any intentionally programmed goals.
3. *Preventing disruption*. If AGI is in the hands of a state or other global actor with uncontested economic dominance, that actor could use AGI to make its dominance persist into the far future.

Another useful angle is to consider what the sources of values change are today, and whether they would necessarily remain operant in a world with AGI. Finnveden et al. (2022) discuss the following sources:

- *Intergroup competition*. As we have seen, warfare and other forms of competition between different states have been a major driver of value change in human history so far. However, we may yet see the emergence of a world government. AGI could even make such an outcome more likely, by providing whoever first develops it with a decisive strategic advantage. Thus if a stable world government arises, competition between states would no longer be a source of values change.
- *Aging and death*. When the leader of an authoritarian regime dies, future leaders may steer things in a different direction, causing values and goals to change over time. Caplan (2008) argues that this problem of succession was the greatest cause of ideological change within the Soviet Union and communist China. In democratic countries too, generational replacement is a source of values change. AGI would not be subject to aging or death, and could therefore continue to pursue values unchanged by this process.
- *Technological or societal changes favouring new values*. In the past, technological or other societal changes have favoured new values. For example, the adoption of agriculture led to a change from egalitarian values to values more accepting of hierarchy. Morris (2015) argues that such values were more adaptive for agricultural societies that relied on more large-scale cooperation and long-term planning. Similarly, because it required more physical strength than other methods of harvesting, plough use encouraged greater gendered division of labor, the effects of which can still be observed today in the form of more unequal gender norms in societies that traditionally relied more heavily on the plough (Alesina, Giuliano, and Nunn 2013). If we reach what Bostrom (2013) calls 'technological maturity', i.e. a level of technological advancement that gives us close to maximum capacity for economic productivity and control over

nature, this source of change would no longer be in play. However, it might stop operating before that, when there are no remaining technological changes sufficiently transformative to overcome the will of a powerful, dominant world government.

- *Internal rebellion*. In the past, coups and revolutions have been a frequent source of values change. In many cases, such regime change has had to rely on support from the military or some other critical state institution. However, if these institutions were no longer reliant on humans, instead being automated by AGI aligned with regime goals, such support would no longer be possible. Moreover, attempts at regime change without the support of key government institutions can also be prevented by AGI.

If AGI-enabled lock-in is feasible, what forms could it take? It could be that the AGI is controlled by some particular state which thereby gains a decisive strategic advantage and becomes able to impose its will globally. In the most extreme case, one could imagine a global totalitarian state where whoever controls the state is able to impose their values around the world. But one could also imagine that a democratic world government only locks something in after extensive public debate and oversight, perhaps also making sure that whatever is locked in will be sensitive to how the will of the people changes in the future. In between these extremes, of course, is a range of different scenarios.

3.2 Lock-in and cultural evolution

It's clear that we want to avoid lock-in by AGI-powered global totalitarianism. But there might be more benign forms of lock-in, such as locking in values chosen in accordance with some democratic process, or locking in procedural elements (i.e. letting the system continue to evolve, but only in accordance with the evolving will of the people, etc., perhaps subject to some further constraints such as human rights, free speech, etc.). Would some lock-in of this kind be desirable? And if so, which particular features would it be desirable to lock in?

When some feature of society gets locked in, that feature is no longer subject to the usual competitive pressures that drive social and cultural change. If we take seriously the idea that it was not our individual intelligence but rather cumulative cultural evolution—the often gradual improvements that have accrued over generations of trial and error—that gave humanity a decisive advantage, we might worry that putting an end to that process risks locking in a suboptimal future. Even supposing that the locked-in feature is currently optimal, it may not remain so as the environment continues to change. This suggests we should be wary of locking in social institutions prematurely. Similarly, today we find at least some of the values of almost any previous historical era to be defective if not outright horrifying. We should expect that future generations will look back on some of our values today in much the same way. For this reason, locking in the specific values we have today might be unwise.

Given that we are bad at intentionally planning and designing effective institutions, Henrich (2015: 331) suggests we ensure that there is sufficient variety and that there are appropriate selection mechanisms, so that different alternatives can compete and evolve. This way, superior arrangements can emerge and spread. We find some support for this idea in the cultural evolution of innovation. Muthukrishna and Henrich (2016) argue that

three key factors behind rates of innovation are sociality, transmission fidelity, and cultural variation. Their starting point is that innovation is more often the result of recombination, gradual improvement, or just pure luck, as opposed to revolutionary leaps by individual geniuses. With this in mind, consider that individuals in populations that are larger and more interconnected will be exposed to a broader range of ideas and practices. If it's not too difficult to learn these new cultural traits (i.e. if transmission fidelity is sufficiently high), some of them will begin spreading through the population. People will use various cues like success and prestige to decide which other people to pay attention to. Assuming that people are at least somewhat able to discern improvements, those improvements will spread to a larger share of the population. For all of this to work, there must be sufficient variation among the ideas and practices that people are exposed to. Too little cultural variation could mean that some superior solution will never be discovered because it can't easily be reached via recombination of current ideas.

MacAskill (2022: 97) similarly suggests we build toward a morally exploratory world. This would mean keeping our options as open as possible, by delaying both large-scale and small-scale lock-in, so as not to risk prematurely ruling out desirable alternatives (Ord 2020: 158). It would also mean a political experimentalism of the kind Henrich gestures at, where people are encouraged to try out different and new ideas. Finally, and crucially, it would require arranging things so that the process of cultural evolution globally guides us toward more desirable arrangements.

What mechanisms could be used to ensure that values, norms, and institutions that are in some appropriate sense better have greater chance to survive and spread? MacAskill suggests it would involve support for free speech so that a broader range of ideas get a fair hearing, relatively free migration so that people can vote with their feet, and international norms or laws preventing any one country from achieving decisive military and economic dominance and unilaterally locking in its goals.

While free speech may encourage a broader range of ideas to be considered, it is hardly a guarantee. Some worry that the global elite may converge on a fairly narrow set of values, practices, and policies. As MacAskill (2022: 96) notes, the relatively small range of global policy responses to the COVID-19 pandemic (e.g. not a single country allowed for human challenge trials of the vaccines developed in early 2020) suggests some such convergence. This trend may only become more pronounced if governance becomes increasingly global. Perhaps even a democratic world government risks leaving too little room for competition.

Thus on this proposal, we should only aim to lock in features that prevent further, undesirable lock-in and guide us toward desirable outcomes we may not have discovered through intentional design. Of course, it may be that future technological developments eventually allow design to outperform cumulative cultural evolution. But we should be wary of prematurely taking ourselves to have reached that point.

3.3 Influencing persistent values

If feasible, AI-enabled lock-in represents the clearest mechanism by which having persistent, predictable influence on the long-term future might be possible. However, cultural evolution suggests there may be other ways of having such influence. Suppose that some set of values provides a sufficiently large competitive advantage in terms of influence over

the long-term future. We should then expect those values to become increasingly prevalent over time, eventually coming to dominate. Plausibly, this is what drove the emergence of large-scale cooperation, as those who were able to more effectively organize into larger units outcompeted others. Are there any other such values that have not yet come to dominate, but will plausibly do so eventually? Hanson (2018) claims that caring explicitly about the long-term future is one such value. He argues that over time, planning and taking action over much longer time frames will become increasingly feasible. Therefore, those who are relatively more inclined to take such actions (rather than actions that are more motivated by short-term concerns) will increasingly have the means to exercise a greater influence on the long-term future, until they eventually come to dominate it.

There are two key steps of this argument that could be questioned. First, why should we expect long-term planning and execution to become increasingly feasible? Second, do we have reason to think that, once long-term planning is possible, taking action with the long-term consequences explicitly in mind will provide a sufficient competitive advantage with respect to long-term influence? Let's consider these in turn. Hanson claims that history so far can be seen as a competition between various kinds of units (organisms, genes, cultures) to control the distant future. So far this has not been very explicit or intentional, because we are not good at planning and taking action over very long time spans. However, he claims that there has been a trend toward more capable long-term planning, and that we should expect this trend to continue. Predators and prey developed the ability to plan for at least part of the duration of a chase. Some animals, like chimpanzees and ravens, are able to plan tool use over several hours (Mulcahy and Call 2006; Kabadayi and Osvath 2017). With farming, humans became able to plan on the scale of a year (e.g. by saving grains to eat in winter and seeds to sow in spring). Today institutions and organizations are able to make some plans on the scale of a few years. This is admittedly a rather small number of examples on which to base our extrapolation, but we can imagine future technological developments that would enable it, like the ones discussed earlier in relation to lock-in. Second, consider now the question of whether explicitly caring for the future will provide a sufficiently large competitive advantage with respect to the long-term future to eventually achieve domination. There is evidence that patience strongly correlates with development, e.g. per capita income and the accumulation of physical capital, human capital, and productivity (Sunde et al. 2022). This suggests that in at least some environments 'long views' may indeed confer a competitive advantage.

If we accept that patient values will eventually come to dominate, what are the practical implications? Hanson suggests that one intervention longtermists might consider is to speed up the arrival of these long views. One might think that, if long views will come to dominate eventually anyway, speeding up their arrival will only have a limited impact on the future. However, as care for the distant future becomes dominant, we will begin investing more in efforts to mitigate extinction risks (assuming that care for the distant future goes together with a belief that continued existence would be good). Therefore, by speeding up the arrival of long views, we reduce the total extinction risk facing us in the future.

How might one work to hasten the arrival of long views? Some possibilities are promoting greater concern for the long-term future in general, making existing cultural units more inclined to care for the long term, or working to make long-term planning more feasible (e.g. by improving our predictive capacities). Further research in this direction might prove useful. What about influencing the broader package of values that go with caring

about the long-term future? There are many different ways of caring about the long-term future. Presumably, not all of these are equally good, and one might therefore think we should work to make better ones more likely.

Hanson further argues, along similar lines, that future entities (whether biological or artificial) will eventually come to directly and explicitly value having as many descendants as possible. So far, we mostly care about descendants in indirect ways. However, again if long-term planning becomes more feasible, those who invest more in taking long-term action will have greater influence on the future. Given such abilities, those who directly plan for having as many long-term descendants as possible will in fact have more long-term descendants than those who don't. This suggests it might be worthwhile to invest further effort in identifying other traits that should reasonably be expected to become dominant in the future. We should then look for a clear way in which the trait provides sufficient long-term advantage, an explanation for why it has not yet come to dominate, and an account of how it may come to do so in the future. This way, we can get a clearer picture of the future landscape of cultural selection pressures that longtermist interventions have to contend with.

4 Conclusion

I have argued that the tools of cultural evolution can inform our thinking about the long-term future. At the most general level, I suggested that for as long as competition between different cultural units remains a relevant force in shaping history, an understanding of cultural selection pressures will be crucial for understanding what the long-term future may look like, and which interventions may be successful. I also claimed that considerations from cultural evolution may support continued experimentation and variation over lock-in and centralization. But the main takeaway I want to convey is that cultural evolution remains an underexplored source of insights relevant to the project of trying to understand and steer the course of the far future. Further work in this direction may well reveal new crucial considerations.

Bibliography

Alesina, A., Giuliano, P., and Nathan N. (2013), 'On the Origins of Gender Roles: Women and the Plough', in *Quarterly Journal of Economics* 128/2: 469–530.

Anthis, J. R. and Paez, E. (2021), 'Moral Circle Expansion: A Promising Strategy to Impact the Far Future' in *Futures* 130: 102756.

Boesch, C. and Boesch, H. (1990), 'Tool Use and Tool Making in Wild Chimpanzees', in *Folia Primatologica* 54/1–2: 86–99.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4/1: 15–31.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

Boyd, R. (2002), 'On Modeling Cognition and Culture: Why Cultural Evolution Does Not Require Replication of Representations', in *Journal of Cognition and Culture* 2/2: 87–112.

Boyd, R. and Richerson, P. J. (1985), *Culture and the Evolutionary Process* (University of Chicago Press).

Boyd, R. and Richerson, P. J. (1996), Why Culture Is Common, But Cultural Evolution Is Rare', in W. G. Runciman, J. M. Smith, and R. I. M. Dunbar (eds.), *Evolution of Social Behaviour Patterns in Primates and Man*, 77–93 (Oxford University Press).

Boyd, R. and Richerson, P. J. (2005), *The Origin and Evolution of Cultures* (Oxford University Press).

Boyd, R. and Richerson, P. J. (2009), 'Culture and the Evolution of Human Cooperation', in *Philosophical Transactions of The Royal Society B* 364: 3281–3288.
Caplan, B. (2008), 'The Totalitarian Threat', in N. Bostrom and M. M. Cirkovic (eds.), *Global Catastrophic Risks*, 504–530 (Oxford University Press).
Cavalli-Sforza, L. L. and Feldman, M. W. (1981), *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton University Press).
Choi, J. and Bowles, S. (2007), 'The Coevolution of Parochial Altruism and War', in *Science* 318/5850: 636–640.
Chudek, M., Muthukrishna, M., and Henrich, J. (2015), 'Cultural Evolution', in D. Buss (ed.), *The Handbook of Evolutionary Psychology Vol. 2*, 749–769 (John Wiley and Sons).
Cotra, A. (2021), 'Why AI Alignment Could Be Hard with Modern Deep Learning', https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/ (Accessed 2025-05-30).
Finnveden, L., Riedel, J., and Shulman, C. (2023), 'AGI and Lock-in,' *Forethought Foundation* https://www.forethought.org/research/agi-and-lock-in (Accessed 2025-05-30).
Greaves, H. and MacAskill, W. (this volume), 'The Case for Strong Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Hanson, R. (2018), 'Long Views Are Coming'. *Overcoming Bias*. https://www.overcomingbias.com/p/long-views-are-cominghtml (Accessed 2025-05-30).
Henrich, J. (2015), *The Secret of Our Success* (Princeton University Press).
Henrich, J. and Boyd, R. (1998), 'The Evolution of Conformist Transmission and the Emergence of Between-Group Differences', in *Evolution and Human Behavior* 19/4: 215–241.
Henrich, J. and Broesch, J. (2011), 'On the Nature of Cultural Transmission Networks: Evidence from Fijian Villages for Adaptive Learning Biases', in *Philosophical Transactions of the Royal Society B* 366/1567:1139–1148.
Henrich, J. and Gil-White, F. J. (2001), 'The Evolution of Prestige: Freely Conferred Deference as a Mechanism for Enhancing the Benefits of Cultural Transmission', in *Evolution and Human Behavior* 22/3: 165–196.
Henrich, J. and Muthukrishna, M. (n.d.), 'What Makes Us Smart?' (unpublished manuscript).
Henrich, N. and Henrich, J. (2007), *Why Humans Cooperate: A Cultural and Evolutionary Explanation* (Oxford University Press).
Heyes, C. (2018), *Cognitive Gadgets* (Belknap Press of Harvard University Press).
Hunt, G. and Gray, R. (2003), 'Diversification and cumulative evolution in New Caledonian crow tool manufacture,' in *Proceedings of the Royal Society B* 270:867–874.
Juniewicz, I. (2022), 'Retrospective on Shall The Religious Inherit The Earth,' *Effective Altruism Forum*, https://forum.effectivealtruism.org/posts/mDkfEjt64DQEtyAmr/retrospective-on-shall-the-religious-inherit-the-earth (Accessed 2025-05-30).
Kabadayi, C. and Osvath, M. (2017), 'Ravens Parallel Great Apes in Flexible Planning for Tool-Use and Bartering', in *Science* 357/6347: 202–204.
Kaufmann, E. (2010), *Shall the Religious Inherit the Earth? Demography and Politics in the Twenty-First Century* (Profile Books).
MacAskill, W. (2022), *What We Owe The Future* (Basic Books).
MacAskill, W., Thomas, T., and Vallinder, A. (2022), 'The Significance, Persistence, and Contingency Framework' in GPI Technical Report No. T1-2022 (Global Priorities Institute, Oxford University).
Mann, J., Stanton, M. A., Patterson, E. M., Bienenstock, E. J., and Singh, L. O. (2012), 'Social Networks Reveal Cultural Behaviour in Tool-Using Dolphins', in *Nature Communications* 3/980.
Mokyr, J. (2016), *A Culture of Growth* (Princeton University Press).
Morris, I. (2014), *War! What Is It Good For? Conflict and the Progress of Civilization from Primates to Robots* (Macmillan).
Morris, I. (2015), *Foragers, Farmers, and Fossil Fuels: How Human Values Evolve* (Princeton University Press).
Mulcahy, N. J. and Call, J. (2006), 'Apes Save Tools for Future Use', in *Science* 312/5776: 1038–1040.
Muthukrishna, M. and Henrich, J. (2016), 'Innovation in the Collective Brain,' in *Philosophical Transactions of the Royal Society B* 371:20150192.
Muthukrishna, M., Morgan, T., and Henrich, J. (2016), 'The When and Who of Social Learning and Conformist Transmission', in *Evolution and Human Behavior* 37: 10–20.
Muthukrishna, M., Shulman, B., Vasilescu, V., and Henrich, J. (2014), 'Sociality Influences Cultural Complexity' in *Proceedings of the Royal Society B* 281: 20132511.
Nakahashi, W., Wakano, J. Y., and Henrich, J. (2012), 'Adaptive Social Learning Strategies in Temporally and Spatially Varying Environments' in *Human Nature* 23: 386–418.

Ngo, R. (2020), 'AGI Safety from First Principles', https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ (Accessed 2025-05-30).
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Scheidel, W. (2019), *Escape from Rome: The Failure of Empire and the Road to Prosperity* (Princeton University Press).
Stark, R. (1996), *The Rise of Christianity* (Princeton University Press).
Sterelny, K. (2021), *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution* (Oxford University Press).
Sunde, U., Dohmen, T., Enke, B., Falk, A., Huffman, D., and Meyerheim, G. (2022), 'Patience and Comparative Development', in *Review of Economic Studies* 89/5: 2806–2840.
Tomasello, M. (2000), 'Two Hypotheses About Primate Cognition' in C. Heyes and L. Huber (eds.) *The Evolution of Cognition*, 165–183 (MIT Press).
Tomasello, M., Kruger, A., and Ratner, H. (1993), 'Cultural learning,' in *Behavioral and Brain Sciences* 16:495–552.
Turchin, P. (2016), *Ultrasociety: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth* (Beresta Books).
Vitousek, P., Mooney, H., Lubchenco, J. and Melillo, J. (1997), 'Human Domination of Earth's Ecosystems,' in *Science* 277/5325:494–499.
Yamamoto, S., Humle, T. and Tanaka, M. (2013), 'Basis for Cumulative Cultural Evolution in Chimpanzees: Social Learning of a More Efficient Tool-Use Technique,' in *PLoS ONE* 8/1: e55768.

PART 3
ETHICAL PRIORITIES

15

The Hinge of History and the Choice between Patient and Urgent Longtermism

Olle Häggström

1 Introduction

Over the past decade it has become increasingly common among scholars working on existential risk and the long-term future of humanity to point to the present time as uniquely and momentously important for the rest of human history. Holden Karnofsky (2021b) speaks of ours as 'the most important century' and 'the one that will initiate, and have the opportunity to shape, a future galaxy-wide civilization'. Here, in a similar spirit, is Max Tegmark:

> It's now, for the first time in the 4.5 billion years history of this planet, that we are at this fork in the road. It's probably going to be within our lifetimes that we're either going to self-destruct or get our act together. (Harris, Goldstein, and Tegmark 2018, 1:16:40 into the recording)

The name, coined by Derek Parfit (2011), that has caught on for this idea is that 'we live during the Hinge of History'. But is that statement true? This was largely taken for granted until the important recent work by William MacAskill (2022a), who addresses this question.[1] After giving the Hinge of History (HoH) concept a (relatively) precise definition, MacAskill offers a probabilistic argument for a 'no' answer: we probably do not live at the HoH, which instead is likely to lie in the far future.

This finding, if it holds up to scrutiny, may have implications for so-called *longtermism*, which we take here to be the view that effects on the very long-term future matter greatly for what we ought to do today. If the HoH is not now but in the future, the best way to promote a good long-term future might not be to take concrete action now, but rather to save and invest our resources until the future true HoH, when we can expect more bang for the buck. Distinguishing between *urgent longtermism*, which recommends taking object-level action now, and *patient longtermism*, which recommends saving and investing now in order to take action later, MacAskill's argument thus speaks in favor of the patient approach.

In the present chapter, I offer a partly opposing viewpoint, defending the idea that it is reasonably likely that we live during the HoH, and also arguing against patient longtermism and in favor of a more urgent approach. The chapter is organized as follows. In section 2 I discuss how to define the hinginess quantity that is maximized at the true HoH, after

[1] MacAskill's paper has a precursor in a tentative blog post (MacAskill 2019a), to which, however, I will not be referring except when required by context, focusing instead on his 2022 paper.

Olle Häggström, *The Hinge of History and the Choice between Patient and Urgent Longtermism* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Olle Häggström 2025.
DOI: 10.1093/9780191979972.003.0015

which, in section 3, I outline MacAskill's so-called base rate argument for why the present time is probably not the HoH. Then, in section 4, I go on to show that his argument is not as strong as it may seem, and I argue that, on the contrary, we have good (albeit far from conclusive) reasons to think that we are in fact living at the HoH. In section 5 I treat the issue of patient vs urgent longtermism, followed in section 6 by some concluding remarks.[2]

2 Defining the Hinge of History

The issue addressed in this section boils down to defining some function H(t)—the hinginess at time t—that captures how pivotal time t is for how well the rest of human history will go. We can then define the HoH as the time t at which H(t) attains its largest value throughout human history, and go on to address the problem of whether the HoH is now. In principle, 'now' might mean 'August 31, 2022' or even be pinned down to a particular time that day, but in the interest of decision relevance or just having a realistic hope of arriving at an interesting answer, it is better to take it to mean something like 'the present decade' or 'the present century'. In connection with his coinage of the HoH concept, Parfit (2011, p. 616) speaks of 'the next few centuries'. It seems reasonable to assume that H(t) is not very volatile and spikey on timescales shorter than decades; otherwise we can replace it by a suitably smoothed approximation $H_{smooth}(t)$, which might for instance be defined as H(t) averaged over the interval $[t, t + \tau]$ for some suitable time window width τ such as $\tau = 100$ years. In the following, unless otherwise specified, I will by 'the HoH is now' mean 'the HoH is in the present century'.

Since we can affect the present and the future but not the past, it can be argued that a more action-relevant definition of HoH would be to consider only times from now and onwards when looking for the maximum of H(t). I will nevertheless stick to the convention of maximizing over *all* times, but will sometimes in what follows be lax about the possibility that the maximum is attained in the past.

To give a definition of H(t) that captures what we want and at the same time is precise enough to satisfy mathematical standards of rigor seems beyond what is presently doable, although there is likely scope for further work in this direction. The definition I will settle on is (almost) the one proposed by MacAskill (2022a) and is somewhat loose around the edges—although not so much so that it prevents a meaningful discussion of whether now is the HoH.

Here's a natural first idea. Assume that we have some utilitarian measure of how well things are going, and let V_∞ denote the total value of the world, from the beginning of prehistory until the end of time. Modeling the trajectory of human history as a stochastic process, this makes V_∞ a random variable. We write F_t as short for a complete description of everything that has happened up until time t, and V_t for the expected value $E[V_\infty|F_t]$ of the total value of the world given its evolution up to time t. Assuming V_∞ to be bounded guarantees existence of this expected value.[3] Now take τ to be a suitable timescale for what we

[2] While the present chapter is largely a reply to MacAskill (2022a), it is worth noting that in his later book (MacAskill 2022b), he looks more favorably upon urgent action, and my impression is that our actual disagreements at the present time are less stark than comes across in the following sections.

[3] This is reasonable if we can ignore everything that happens outside our future light-cone and after the heat death of the universe, but might, depending in part on unsettled cosmological issues, turn out untenable,

take to be the duration of 'now', such as $\tau = 100$ years. By the tower property of conditional expectation (Williams 1991), we have $E[V_{t+\tau}|F_t]=V_t$. In other words, given all that has happened up to time t, the expected change over the time interval $[t, t + \tau]$ in the conditional expectation of V_∞ is precisely 0. Then the conditional variance $Var[V_{t+\tau}|F_t]$ suggests itself as a candidate for a measure of how much is at stake at time t.

However, the problem with taking H(t) to be $Var[V_{t+\tau}|F_t]$, from the point of view of the policy-relevant issue of patient vs urgent longtermism treated by MacAskill (2022a) and in section 5 below, is that it fails to take into account the aspect of human agency. Imagine (counterfactually) that in the year 1200, a comet large enough to wipe out all large land-living animals including *Homo sapiens* were on collision course to impact the Earth in 1250, with the proviso that its passages near Jupiter and through the asteroid belt would involve chaotic dynamics making hit-or-miss a 50-50 chance issue in our world model. That would make $Var[V_{1300}|F_{1200}]$ very large and the 13th century extremely hingey—perhaps enough to make it the true HoH—and yet it would not constitute a reason for 13th-century decision-makers to take immediate longtermist action, because this was so long before reaching a technological level enabling a detection-and-deflection program that there was simply nothing anyone could do to affect the outcome.

So from the policy-relevance perspective, it would be desirable to pin down how much of $Var[V_{t+\tau}|F_t]$ emanates from human decision-making between times t and $t + \tau$. Instead of going down this path, however, I will follow more closely that of MacAskill (2022a), which is more explicitly tied to the policy issue at hand. He looks at how much a philanthropist at time t can improve V_t per amount of resources spent, and takes H(t) to be the optimal such ratio. The basic logic is simple: if now is time t, if $t' > t$, and if $H(t') > H(t)$, then this suggests saving the resources until time t' and using them then as a way to achieve a greater improvement in V_t compared to using them now. (Of course, issues involving interest rates and financial risk greatly complicate this basic logic; see Trammell (2021) for an in-depth treatment.)

This definition of H(t) is a stepping stone towards defining the HoH, and I do it slightly differently from MacAskill, who looks at influentialness of individual people rather than hinginess of times, but the reasoning is essentially the same. He does, however, bring up some caveats and elaborations, two of which merit repeating here: one about (i) the definition of 'spend', and the other about (ii) what sort of spending counts as achievable. As regards (i), spending here only refers to direct actions towards improving V_t, as opposed to, e.g., investments or movement building aimed at attaining more resources for taking such action in the future. The possibility of gray areas (treated in more detail by Cotton-Barratt 2020, and Mogensen 2022) in this distinction does not preclude having a principled discussion about hinginess and the HoH.

Concerning (ii), MacAskill emphasizes that we should only consider those actions as available at time t for which there is information available to philanthropists and other decision-makers at the time that gives plausible reasons to believe that the given action actually does improve V_t. This means that if I had been a citizen of Vienna in 1889 with an opportunity to kill baby Hitler, and if such an action would have had a great positive impact on political and other developments in the 20th century, that would still not

causing deep difficulties that however fall outside of the ambitions of the present chapter (see, e.g., Bostrom 2011; Askell 2018).

count in this context, because at the time I would have had no way of knowing that the murder would be a good thing. I will not here address the topic of cluelessness (Greaves 2016), which is the idea that perhaps the long-term effects of all our actions are similarly intractable, except to say that I am sympathetic to the stance of Greaves (2020) and Schubert (2022) that while the concept does bring up interesting concerns, there are still plenty of things we can do that make sense from the longtermist perspective of improving V_t.

Finally, when it comes to defining the HoH, I will do the straightforward thing of taking it to be the time t at which H(t) is maximized. This may at first look like it is in full agreement with MacAskill's understanding of the HoH-is-now hypothesis to mean that

> we are among the very most influential people ever, out of a truly astronomical number of people who will ever live (MacAskill 2022a: 339),

but it isn't, due to the proviso 'out of a truly astronomical . . .'. If human history ends in extinction today, then the total number of humans who ever lived will stop at something like 10^{11}, which is not 'a truly astronomical number', so that on MacAskill's definition there will be no HoH. MacAskill is explicit about this proviso and has a clear reason for his choice (which I will come back to in section 3). Still, the proviso is somewhat unnatural (and Mogensen (2022: 3), goes as far as saying that it 'distorts the issue'), so I will drop it here. If H(t) is higher now than ever before, and we fail—in the above-quoted words by Tegmark—to 'get our act together' but instead cause our near-term extinction, then I will say, contra MacAskill, that we did now live at the HoH, but we blew it. (Let's not!)

3 MacAskill's base rate argument against the Hinge of History being now

One of the most common arguments for longtermism involves back-of-the-envelope calculations in the style of Bostrom (2013) to show how much potential value there is in the future provided we avoid near-term extinction. Conservative assumptions about a kind of sustainable business-as-usual civilization refraining from space colonization lead Bostrom to suggest the potential for 10^{16} future human lives of normal (10^2 years) duration, while with interstellar and intergalactic space colonization 10^{32} lives may be feasible, and in case of a workable mind-uploading technology, perhaps 10^{52}. The point here is not the exact orders of magnitude (which may well be somewhat off) but rather how very large these numbers are, in particular compared to numbers such as 10^{10} that a naive short-termist thinker (or a proponent of so-called person-affecting views; see, e.g., Greaves 2017) might otherwise be tempted to use to represent how many human lives are at stake in connection with existential risk.

For his main argument against the HoH-being-now hypothesis, MacAskill (2022a) combines such population estimates with another Bostrom idea, namely the formalization and quantification of the Copernican and mediocrity principles that has become known as the Self-Sampling Assumption (SSA), which Bostrom (2002) defines as follows:

> One should reason as if one were a random sample from the set of all observers in one's reference class. (p. 57)

The choice of 'reference class' may here depend on background information and the particular application at hand, and is often open to debate; in my case, the reference class could be, e.g., the set of all present-day Swedish-born professors of mathematics, or of all human beings throughout past and future history, or of all sentient beings who ever existed or will exist in the entire universe.

Now assume that one or the other of the future scenarios underlying Bostrom's population calculations is roughly what will happen, take $n = 10^{10}$ (slightly above the world population as of today), and let N be the total number of humans throughout past and future history. MacAskill notes that the ratio n/N will then be a very small number; exactly how small is less interesting, but he speaks about 'one in a million trillion' (MacAskill 2022a: 341), so for concreteness let's say $n/N = 10^{-18}$. Imagine ranking the N humans that will ever live according to the peak value of hinginess H(t) that they experience during their lives (with some arbitrary tie-breaking convention), and note that applying SSA with reference class consisting of all N humans gives me probability $n/N = 10^{-18}$ of being among the top n people on this ranking. If the HoH is now, then clearly I will be among these top n, so the probability that the HoH is now can be at most 10^{-18}.

MacAskill doesn't mean this to be an upper bound of the probability, given the best of our knowledge, that we are living at the HoH. Rather, it is the *base rate*, to be used as a Bayesian prior, and then modified by conditioning on all the evidence for or against living at the HoH that we get when we look around at the world in all its hinginess. Since 10^{-18} is such an extremely small number, the statement that 'we live during the HoH' can be seen as quite an extraordinary claim, suitable for an application of the famous Carl Sagan maxim that 'extraordinary claims require extraordinary evidence'. And while MacAskill does admit that we do have evidence for living at the HoH, he holds evidence to be insufficiently extraordinary to bring the probability from 10^{-18} all the way up to (say) double-digit percentages. The probability that we live at the HoH is therefore small.

Note, however, that MacAskill's argument requires the premise of a large future à la Bostrom: it says nothing, for instance, to rule out the possibility of H(t) taking its maximum value now, promptly followed by an existential catastrophe wiping out humanity. Now we understand the inclusion of 'a truly astronomical number' in his preferred definition of HoH: it allows him to state his main conclusion more crisply as a claim about a probability rather than a conditional probability. The crispness comes, however, at the expense of potential misleadingness to cursory readers. This is not, however, among my main issues—to be treated in the next section—with MacAskill's argument.

4 Counterarguments

I will offer two separate counterarguments to MacAskill's application of SSA for arguing against the hypothesis of the HoH being now. The first will be a general caution against giving too much credence to conclusions derived from the SSA principle and related anthropic arguments, and the second will be an observation that a prior probability of $1–10^{-18}$ is not as overwhelmingly unyielding to contrary evidence as first meets the eye.

Regarding how much credence to attach to SSA-based reasoning, we should first note that the principle is nowhere near as well established and widely tested as other epistemic principles such as Bayesian updating in the light of new evidence or Ockham's razor. While

SSA-like ideas go back a bit further than Bostrom (2002)—see, e.g., Gott (1993) and Leslie (1998)—it remains the case that SSA and related principles have undergone very little practical testing. Applications of SSA are limited mainly to simple thought experiments and large-scale (often cosmological) questions whose answers remain unknown, with very little in between. If SSA amounted to a fairly obvious and uncontroversial claim, this in itself would perhaps not be a big problem, but that is not the case. On the contrary, there are several serious challenges regarding its validity, and while I do think SSA is an excellent idea that merits our attention, I also think that those challenges need to be taken seriously, including the following.

An obvious problem with SSA is the vagueness of what constitutes the correct reference class, but even if that can be settled, we must recognize that the idea that we can view ourselves as chosen from this class according to uniform distribution is a non-innocent probabilistic model assumption (Häggström 2007). This model assumption needs better grounding, especially since no mechanism for it has been proposed. Such mechanisms would seem to require dualistic ideas along the lines of God picking each of us from some reservoir of souls and assigning us to bodies chosen according to uniform distribution in the appropriate reference class—so, naturally, Bostrom and other defenders of SSA instead fall back on *the principle of indifference*, which was important in the early (pre-Kolmogorovian) development of probability theory, and states that in finite settings without obvious asymmetries between outcomes, equal credence should be assigned to all outcomes. Indeed, in the basic thought experiments used by Bostrom (2002), the reference classes have such far-reaching symmetries that it becomes difficult to defend a violation of the principle of indifference. In real-world applications such as MacAskill's, there is, however, an abundance of asymmetries. Why insist on uniform distribution on the class of observers, when probabilities could well depend on, say, observers' longevity,[4] their earliness in history, the energy turnover of their brains, their intelligence, or who knows what?

Another concern regarding the SSA is the mathematical impossibility, due to the fact that no probability measure on the set of natural numbers exists which assigns equal probability to all elements, of applying SSA in an infinite universe with infinite reference classes. Furthermore, in the finite case, applications of SSA tend to land in unpalatable consequences involving Doomsday arguments (Leslie 1998; Bostrom 2002; Häggström 2016; Thomas 2021), simulation hypotheses (Bostrom 2003), and Boltzmann brains (Carroll 2021). While I do have some sympathy for the intellectually courageous practice of biting the bullet and accepting wild consequences of otherwise reasonable-looking model assumptions, it still makes sense to take such consequences as additional evidence against the model assumptions if these are already seen to stand on shaky grounds.

All of this suggests that while an SSA approach to a problem like whether we live at the HoH provides a valuable perspective, the results of such an analysis need to be taken with a grain of salt. This is especially true when, as in the present case, a hypothesis which does have evidence pointing in its favor is assigned a very small probability in the SSA analysis.

Let me nevertheless, for the sake of argument and for concreteness, assume that MacAskill's analysis as outlined in section 3 yields an appropriate prior for the HoH problem, and take the prior probability that we do *not* live at the HoH to be the seemingly

[4] In fact, Bostrom (2002) himself goes on, via the concept of *observer moments*, to advocate a refinement of SSA where observers get probabilities proportional to their longevity.

overwhelming $1-10^{-18}$. What will be the consequences of this when we update the prior by conditioning on further information about our present situation?

It may be tempting to think of the prior probability $1-10^{-18}$ of not living at the HoH as so overwhelming, and of the extraordinariness of the evidence needed to overcome it as so great, that pretty much no evidence short of a miracle will do. That would be a mistake, however, as in the standard statistical setting with independent and identically distributed likelihood observations, ratios increase (or decrease, depending on one's point of view) exponentially with sample size.[5] Shaving an order of magnitude off an odds ratio takes a constant, and in many cases not particularly large, number of observations: to overcome 100 to 1 odds enough to equalize takes roughly just twice as large a sample as for overcoming 10 to 1 odds, and overcoming 10^{18} to 1 odds takes a sample size that exceeds the original one by a mere factor 18.

As an example, consider coin tossing. Suppose that we have a coin whose heads-probability q is strictly speaking unknown, but whose fairness ($q = 0.5$) we have such strong prior reason to believe in that when we set up the *a priori* distribution to reflect these prior reasons fairly, we assign probability $1-10^{-18}$ to the event that $q = 0.5$. The remaining probability 10^{-18} is smeared out uniformly on the interval where q ranges from 0 to 1. How many coin tosses with 75% of them heads would it take overcome such a dogmatic-looking prior? Perhaps surprisingly, the answer is not astronomical. A straightforward calculation shows that 1,000 tosses with 750 heads vs 250 tails would be enough to totally turn the tables and produce a posterior distribution that places probability more than $1\text{-}10^{-35}$ on the coin being biased with $q > 0.5$.

This shows that overcoming an *a priori* distribution as heavily slanted as MacAskill's against a given proposition need not be as big a deal as it might first appear. Quantifying the available evidence that our time is the true HoH is, however, a delicate matter that I will not be able to resolve here, and a skeptic might argue that while we do have some evidence for living at the HoH, it is doubtful whether it amounts to 1,000 independent bits of information. Nevertheless, I will point to some concrete reasons why it is not unreasonable to think it may be enough overcome the 10^{18} to 1 odds that MacAskill stacks against it.

Consider for instance the fact that we live within the 100-year period immediately following our first setting foot on a planetary body other than our home planet. It is hardly implausible to suggest that such an event might be special enough for what happens the century following to determine the rough structure of our subsequent colonization of the universe. This is obviously not a water-tight argument for the HoH necessarily happening within a century from that pioneering first step of space exploration, but assigning a prior probability of 1 in 10 to such a concurrence does not strike me as crazy. If we can do that, then the success of the Apollo program allows us to immediately cancel 17 of the 18 orders of magnitude contained in the odds ratio in MacAskill's prior. At that point, finding further evidence to overcome one more order of magnitude (or a few more, to take the probability of HoH being now past the even odds equilibrium and towards domination of the posterior distribution) no longer seems like such a daunting task.

[5] The mistake is understandable in view of today's scientific practice of designing experiments with small sample sizes meant to produce p-values around 0.05 or 0.001 being so dominant that we tend to forget about the exponential behavior of likelihood ratios as sample size increases.

Similarly, we may consider the evolution of intelligence. Humanity's astonishing trajectory from being a relatively unremarkable species a million years ago towards world domination has had relatively little to do with, say, our muscular strength or our physical endurance: what sets us apart from other species is almost entirely our intelligence. The step of beginning to outsource this unique resource to machines therefore seems like at least as significant a transition in human history as those first steps on the Moon, so assigning a probability of 1 in 10 to the HoH happening in the same century as the breakthrough in artificial intelligence (AI) seems reasonable, and that gives us another reason to cut 17 of the 18 orders of magnitude contained in MacAskill's odds against the HoH being now. (On the other hand, treating the breakthroughs in space exploration and in AI as statistically independent events seems wrong, and we cannot naively stack those 17 orders of magnitude gained from the latter on top of those 17 from the former to arrive at the probability of our time being the HoH being an impressive $1-10^{18-17-17}=1-10^{-16}$.)

These examples are in the spirit of Shulman (2019), who holds forth the Moon landing and a few other similarly unique aspects of our time as indicative of the HoH. However, a skeptic such as MacAskill (2019b) may respond that these must be compared to other potentially great transitions in the future in order to carry much credence. The reason we think of the dawning of space exploration and of AI as so significant to human history might simply be myopia: we are smack in the middle of these transitions, and it's only natural that we overestimate the importance of the here and now. We might, at some time in the far future, look back on these events and view them as relatively small steps compared to, say, the initiation of a controlled merger of the Andromeda and Milky Way galaxies, or the rollout of a technology for harvesting energy from false vacuum decay, or something entirely beyond our present imagination. This is possible, but seems to me at most moderately likely.

We could also look for more generic clues to our place in history. Based on an analysis by Hanson (2009), Karnofsky (2021a) looks at global economic output (operationalized as inflation-adjusted global GDP, but the details do not matter much), and finds this:

> Let's say the world economy is currently getting 2% bigger each year. This implies that the economy would be doubling in size about every 35 years. If this holds up, then 8200 years from now, the economy would be about $3{\cdot}10^{70}$ times its current size. There are likely fewer than 10^{70} atoms in our galaxy, which we would not be able to travel beyond within the 8200-year time frame. So if the economy were $3{\cdot}10^{70}$ times as big as today's, and could only make use of 10^{70} (or fewer) atoms, we'd need to be sustaining multiple economies as big as today's entire world economy *per atom*. (Emphasis in original)

The last conclusion can be treated as a *reductio*, meaning that until the time we begin colonizing other galaxies, there will not be as many as 82 future centuries exceeding ours in relative GDP growth. Similar exercises with GDP replaced by energy consumption or population of flesh-and-blood humans lead to similar results; see Murphy (2011) and Abell (1982), respectively, with the latter culminating in the vivid image of a sustained 2% annual population increase leading within 5,300 years to 'a great sphere of humanity 150 light years in radius [which] would be expanding at its surface at the speed of light' (p. 594). Sticking for concreteness to Karnofsky's calculation, one might object that a greater number of centuries with such GDP increase can be achieved if GDP is allowed to oscillate, but the argument holds up against this objection by replacing momentary GDP by all-time-high GDP

in the analysis. Also, if we replace 10^{70} with the upper bound of 10^{82} on the number of atoms in the (reachable) universe, we get at most 95 future centuries of present-day GDP growth or more, even without the restriction of staying in the Milky Way. Plausibly, times with an unusual amount of change also come with unusual hinginess, so it seems reasonable to assume that the probability of the HoH happening in one of the 95 centuries exhibiting the most drastic economic growth is at least 1/2. Conditioning on the fact that we live in such a century, we arrive at a probability of $1/(2{\cdot}95) = 0.0053$ that the HoH is in the present century, thereby overcoming nearly 16 of the 18 orders of magnitude in MacAskill's prior.

Another circumstance that points towards the probability of our living at the HoH being far larger than 10^{-18} is how very early in human history we find ourselves in relation to the hugeness of the Bostromian futures that do most of the work in MacAskill's argument: it is not unreasonable to think that the long-term fate of humanity is disproportionally determined by what happens very early. In the comments section to MacAskill (2019a), such considerations led Ord (2019) to suggest a Laplacean $1/t^2$ prior for the timing of the HoH, and 'a prior chance of [the present being the] HoH of about 5% or 2.5%'.

Relatedly, our living at a time when human population counts in billions rather than the quadrillions or more, as in those Bostromian futures (aided, perhaps, by space colonization and/or mind uploading), presumably makes it easier for individuals and small groups to have momentous influence on the remainder of human history, again pointing towards our time exhibiting a greater-than-usual hinginess and thereby a larger probability of being the HoH. Rather than plunging into (shaky) attempts to quantify the evidential value of these two circumstances, let me move on to yet another one, namely the existential risk that is gestured at in statements like the Tegmark quote in section 1.

Assume for the moment that during the present century we run a 10% risk of a catastrophe resulting in the extinction of the human species. How many similarly dangerous (or worse) centuries can there be? For there to be n such centuries, we need to survive the first n-1 of them, which has probability at most $(1–0.1)^{n-1}$. For instance, taking n = 100 yields probability at most $(1–0.1)^{99} = 0.00003$ of having at least 50 centuries with such a high extinction risk. It seems plausible to postulate at least even odds for the HoH to happen during one of the 100 most dangerous centuries in terms of extinction risk, which would put a lower bound on the probability of the HoH happening during the present century at just under 0.5/100 = 0.005. Again, most of the orders of magnitude in MacAskill's odds against the HoH being now are wiped out.

But how realistic is it to postulate such a 10% risk of extinction catastrophe? For natural risks such as supervolcanoes and asteroid impacts, we do have a reasonable grip on estimating the risk level, and on timescales of a century they are small. Estimating anthropogenic risk is harder, as circumstances here are so new and rapidly changing that no scientific consensus exists that pinpoints even the appropriate order of magnitude of the total risk, but this is where the risk must come from in order to come anywhere near the suggested 10%. The magnum opus on existential risk (x-risk) by Ord (2020) estimates, along with suitable declarations of epistemic modesty, the x-risk probability during the next 100 years as 1/6. This, however, is not immediately applicable to the calculation in the preceding paragraph, because the concept of x-risk does not quite equal that of extinction risk: Ord follows Bostrom's (2013) standard definition of x-risk which besides extinction risk includes other events on a similar level of badness in terms of losing humanity's potential for future flourishing.

To bridge the gap between x-risk and extinction risk in our calculation we can choose either of two approaches. One is to note that on a centennial timescale, and judging from Ord (2020) along with other works on existential risk including Bostrom and Cirkovic (2008), Pamlin and Armstrong (2015), Häggström (2016), and Yudkowsky (2022), extinction risk seems to be the main and most likely example of x-risk, so we could take Ord's 1/6 and round it down to something like 1/10 to get a plausible extinction risk probability. The other approach is to apply the calculation to x-risk rather than extinction risk, which, however, gives rise to the complication that unlike extinction events, existential catastrophes can in principle happen more than once: the first time around, humanity survives but loses most of its potential future value; the second time, most of the remaining potential value is lost, and so on. This can arguably be fixed by noting that the HoH will likely happen before (or during) the first existential catastrophe, because after that most of the potential value is already out of reach, so it makes sense to consider only those centuries that predate the first existential catastrophe, and with that restriction the calculation goes through as for extinction risk.

As mentioned, there is considerable scope for debate regarding Ord's (2020) assignment of probability 1/6 to x-risk during the next 100 years, but it does not strike me as an obvious overestimate. Regarding the specific risk source that he gives the largest probability—unaligned AI, which accounts for probability 1/10—it seems to me that on the contrary, the best state-of-the-art judgments point towards, if anything, the risk being even greater. See, e.g., Ngo (2020), Häggström (2021), Carlsmith (2022), Christiano (2022), and Cotra (2022a; 2022b) for recent discussions of the gravity of AI x-risk, and Yudkowsky (2022) for a particularly alarming but not obviously erroneous account of the difficulty of managing that risk.[6,7]

Still, in view of the large uncertainties, it is not clear how much weight to give x-risk in the HoH discussion. As a stand-alone argument for our living at the HoH, I find it relatively strong, but as an objection to MacAskill's (2022a) base rate argument it is open to the counter that Ord's 1/6 estimate has not taken the base rate argument into account and might therefore need to be adjusted down.

Be that as it may, we have in the above seen a wide array of circumstances that make our time unusual in a way that should boost our credence in it being the HoH. I have not provided any mathematical model for integrating this evidence and deriving a Bayesian posterior, which would be a daunting task and probably not achieve much beyond a false air of precision, but I can offer my subjective judgment of where the evidence seems to take us: a reasonably high probability of our living at the HoH, most likely a double-digit percentage number, and plausibly even the majority of the probability mass.

However, as good Bayesians we cannot be content with looking for evidence in favor of a given proposition. Rather, we should aspire to take into account all relevant information, including that which points in the opposite direction. MacAskill (2022a) suggests the following circumstance as evidence against the HoH being now: our ability to affect the future for the better depends on our knowledge about how the world works, and this (collective) knowledge has improved in the past and can be expected to improve further in the

[6] Well-argued dissenting views are rare, but Hanson (2019; 2023) is among the better.

[7] See also Finnveden, Riedel, and Shulman (2022) for other AI-related reasons (beyond AI x-risk) for our time being the HoH.

future, pushing our ability and therefore the level of hinginess in the same direction, thus improving the chance that the HoH is in the future rather than now. This is an interesting argument, but it is unclear whether it works, because it may be that the most relevant quantity, given the definition of hinginess H(t) recalled in section 2, is not the absolute level of knowledge but rather the knowledge gap between on one hand philanthropists aiming to increase V_t, and on the other hand decision-makers in general; the absolute level increasing over time need not imply that gap does. I will return briefly to this issue in section 5.

Relatedly, MacAskill suggests that the anomalous GDP growth (along with other rapid societal changes) discussed above might not tend to increase hinginess but rather to decrease it, because it might be easier for V_t-maximizing philanthropists to influence the world in periods with little change when they are in a better position to predict the consequences of their actions, and are therefore better able to choose the best actions. This could be, but one can also reason in the opposite direction and argue that in times of technological and societal stasis it is very difficult to have much impact, and that the best opportunities for impact are when society opens up to change. One can argue back and forth over which of these effects is the stronger. I believe more in the latter, but even if we grant MacAskill even odds on this issue, the effect on the probability distribution of hinginess of our present turbulent time (relative to others) will mostly be to push probability from the middle and towards the extremes, resulting in an increased probability that we live at the HoH.

5 In defense of urgent longtermism

Recall from section 1 the distinction between urgent and patient longtermism, and how the choice between them relates to the HoH issue. The arguments offered in the previous section for the likelihood of our living at the HoH point somewhat in the direction of favoring an urgent approach, while still falling far short of settling the debate, mainly for two reasons. First, while I have argued that it is reasonably likely that the HoH is now, this does not entirely eliminate the possibility that the HoH lies in the (possibly far) future. Second, even if the HoH is now, there could be other times in the future that also exhibit high hinginess, and by saving in financially cunning ways that give good return on investment and delaying action to those later times, we might create more expected value. So more can be said on the choice of patient vs urgent longtermism, and the following are a few thoughts on this matter.

Consider for concreteness a scenario where our present century is fairly hingey, but not quite as hingey as the true HoH that takes place a million years from now. The choice of timescale here may look like a straw man of MacAskill's (2022a) position, but this sort of very large timescale is in fact needed in his base rate argument for a very small prior probability of the HoH being now. For further concreteness, add the assumption that from now on and over the next million years, we are able to raise $1bn (one billion dollars) per year for longtermist purposes. Ignoring for the moment the issue of interest rates, we may note the asymmetry that, among the full $\$10^6$bn raised over this period, only at most $100bn can be spent during the present century, while potentially the entire $\$10^6$bn is available at the HoH at the end of this period. Thus, in this idealized situation, the patient vs urgent longtermism issue boils down to whether to spend $100bn this century and the remaining $999,900bn (i.e., about ten times present-day annual GDP) during the HoH a million years hence, or to save the entire amount until then.

It is by now a well-established wisdom in the effective altruism community that for individual donors, the phenomenon of diminishing marginal returns is usually negligible, so the effect of spending on a given cause can be seen as linear and one might as well concentrate all one's spending on one single cause, but that on larger scales diminishing returns may be more substantial (see, e.g., Kuhn 2014; Snowden 2019). Whether in the situation outlined here $100bn in the present century has more impact than the additional 0.01% of longtermist funds at the distant HoH is of course an open question, but it seems at least plausible that diminishing returns would cause the urgent spending to be the more impactful option.

Next, consider the effect of interest rates. Even a modest real interest rate of 1% per year would mean a doubling of capital in 70 years, and a multiplication by 2.7 each century. So the value a million years down the road of the $100bn raised in the present century will be $100bn times $2.7^{10,000}$, which means, with a = 2.7, that the present century's fraction of that raised in the next million years will be

$$1/(1+a^{-1}+a^{-2}+\ldots+a^{-10,000}) = (1-a^{-1})/(1-a^{-10,001}) = 0.63.$$

This is a very different story from that obtained with a zero interest rate, but $100bn·$2.7^{10,000}$ exceeds 10^{4000}—which in turn far exceeds the number of elementary particles in the visible universe or pretty much any (non-combinatorial) number that arises in our world, so we see that the result is nonsensical for essentially the same reason as in the GDP and population growth examples in section 4 (i.e., sustained exponential growth quickly becomes incompatible with finiteness of the speed of light). In conclusion, working with constant interest rates over such enormous time spans is infeasible. Grabbing for straws, perhaps something like hyperbolic discounting (Ainslie 2001) could come to the rescue, but I think the grim truth is that we have at present no way of making sense of the value of financial assets over millennial or longer timescales.

We could, however, point to some more concrete issues concerning such extremely long-term investments. One is that the concept of money is just a few millennia old, and modern capitalism only a couple of centuries, so that saving over timescales longer than that implies a faith in the longevity of these institutions that appears unwarranted. Who knows what kinds of governmental confiscations, wars, revolutions, and civilizational collapses the next million years may have in store for us?

On the more positive side, consider moral progress, of which recent centuries have seen plenty, as MacAskill (2022a) points out in connection with the argument about the trend towards improved knowledge discussed in section 4 above:

> In 1600 [it was] believed that women and people of other races and religions are of lesser moral standing than European Christian men. Intense social hierarchy, inequality and slavery were regarded as the natural and just way of things. Homosexuality and premarital sex were regarded as deeply immoral. The idea of liberalism had not been developed. Torture was commonplace and celebrated, as were cruel punishment and violence against heretics. (347)

These views and practices are now understood as morally wrong, and MacAskill goes on to stress that the trend continues and that in the last 50–100 years 'rights for women,

minorities and people of all sexual identities have been progressively secured' (348). His point here (echoing the famous final paragraph of Parfit 1984) is that this development towards better morality can be expected to continue in the future, and I agree, but I think the consequence of this may go beyond what MacAskill suggests about future longtermists being better equipped to know what to do. Consider the following.

Assuming that saving in longtermist funds does not influence GDP growth in any particular direction, the point of these savings is to exert influence on how society in the future will use their resources, by earmarking some of these for longtermist goals. If human morality continues to improve, and if V_t-maximization is a largely correct stance (otherwise the whole patient vs urgent longtermism issue seems moot, and it is probably not a good idea to constrain future generations to act unethically), then we can expect society as a whole to eventually adopt these ethical views and act upon them, thereby removing the need for us to lock up resources in longtermist funds in order to constrain future people to act on these views.

In practice, yet another concern one may have regarding patient longtermism is the following. While near-term and long-term goals for society are often fairly well aligned, there will be cases when they come somewhat apart, and it is often difficult in such cases to convince individuals and society as a whole to sufficiently prioritize the long term (as evidenced by, e.g., climate debate). It is likely that this difficulty is exacerbated if the long-term action is to put money in the bank rather than, say, taking concrete action against some x-risk.

The discussion so far points mostly in the direction of preferring urgent longtermism, but it is interesting to stop and consider how general the arguments are. If they suggest that taking concrete action is better than patiently saving now in the 21st century, will the conclusion be the same in, say, the 22nd century, or the 24th? Are the arguments in fact so general as to suggest that we should *never* adopt a patient longtermist stance? Well, while some of the arguments are fairly general, I'd say the largest cornerstone in the case for urgent longtermism presented so far consists of the evidence discussed in section 4 for the unusually large hinginess of the present century. If this (apparent) hinginess level goes down in later centuries, the case for urgent approaches will then become weaker.

One can turn this question around and ask whether, for someone who accepts MacAskill's base rate argument, we can ever expect to see evidence of being at the HoH that is so much stronger than today's already quite remarkable evidence in this direction that it overcomes the low base rate and recommends urgent action? Maybe, but then again maybe not, and it would be sad to see longtermist funds remaining untouched forever.

6 Concluding remarks

Despite all the evidence presented in section 4 for our living at the HoH, the question for whether or not that is the case remains wide open. Likewise, the discussion in section 5 hardly settles the issue of patient vs urgent longtermism. While I do think the idea of setting up longtermist funds with resources to be harvested in a million years can (at least for the time being) be written off as impractical and overly fanciful, there is the much more modest idea that the 22nd or the 24th century might be even hingier than the 21st, and that a good longtermist move would be to save resources until then. The arguments in the present chapter do relatively little to invalidate such a moderate version of patient longtermism. On

the other hand, such a view receives hardly any support from MacAskill's base rate argument, which needs much longer timescales to produce the seemingly overwhelming prior odds that he holds forth against our living at the HoH.

Regarding the choice between urgent action and this more moderately patient longtermism, there is the following basic intuition. While it remains a highly open question which of these would be recommended by a well-informed utilitarian-style calculation, there is an asymmetry in that we pretty much know that the level of existential risk in the present century is alarmingly high (in particular, it has become increasingly clear in the last couple of decades that AI existential risk is substantial). If we fail to take concrete action in mitigating this risk, we may face an existential catastrophe that cannot be repaired no matter the amount of resources put aside for longtermist work in later centuries. I think this intuition makes sense, and it is lent some support by the arguments in the present chapter that urgent longtermism is at least not obviously wrong.[8]

References

Abell, G. (1982), *Exploration of the Universe*, 4th edition (Thomson Learning).
Ainslie, G. (2001), *Breakdown of Will* (Cambridge University Press).
Askell, A. (2018), *Pareto Principles in Infinite Ethics*, PhD thesis, New York University.
Bostrom, N. (2002), *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge).
Bostrom, N. (2003), 'Are You Living in a Computer Simulation?', in *Philosophical Quarterly* 53: 243–255.
Bostrom, N. (2011), 'Infinite Ethics', in *Analysis and Metaphysics* 10: 9–59.
Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4: 15–31.
Bostrom, N. and Cirkovic, M. (2008), *Global Catastrophic Risks* (Oxford University Press).
Carlsmith, J. (2022), 'Is Power-Seeking AI an Existential Risk?', https://arxiv.org/abs/2206.13353 (accessed August 15, 2022).
Carroll, S. M. (2021), 'Why Boltzmann Brains Are Bad', in S. Dasgupta, R. Dotan and B. Weslake (eds.), *Current Controversies in Philosophy of Science* (Routledge), 7–20.
Christiano, P. (2022), 'Where I Agree and Disagree with Eliezer', *LessWrong*, 19 June, https://www.lesswrong.com/posts/CoZhXrhpQxpy9xw9y/where-i-agree-and-disagree-with-eliezer
Cotra, A. (2022a), 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover', *LessWrong*, 18 July, https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to
Cotra, A. (2022b), 'Two-Year Update on My Personal AI Timelines', *LessWrong*, 2 August, https://www.lesswrong.com/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines
Cotton-Barratt, O. (2020), 'Patient vs Urgent Longtermism Has Little Direct Bearing on Giving Now vs Later', *Effective Altruism Forum*, 9 December, https://forum.effectivealtruism.org/posts/Eh7c9NhGynF4EiX3u/patient-vs-urgent-longtermism-has-little-direct-bearing-on
Finnveden, L., Riedel, J., and Shulman, C. (2022), 'Artificial General Intelligence and Lock-In', https://docs.google.com/document/d/1mkLFhxixWdT5peJHq4rfFzq4QbHyfZtANH1nou68q88 (accessed August 15, 2022).
Gott, R. (1993), 'Implications of the Copernican Principle for Our Future Prospects', in *Nature* 363: 315–319.
Greaves, H. (2016), 'Cluelessness', in *Proceedings of the Aristotelian Society* 116: 311–339.
Greaves, H. (2017), 'Population Axiology', in *Philosophy Compass* 12: e12442.
Greaves, H. (2020), 'Evidence, Cluelessness and the Long Term', talk at the virtual *Effective Altruism Student Summit*, 24–25 October, with transcript at the *Effective Altruism Forum*, 1 November.
Häggström, O. (2007), 'Uniform Distribution Is a Model Assumption', http://www.math.chalmers.se/~olleh/reply_to_Dembski.pdf(accessed August 15, 2022).

[8] I am grateful to Gustav Alexandrie, Björn Bengtsson, Hilary Greaves, Erich Grunewald, Julia Jansson, Stefan Schubert, Christopher Star, Markus Stoor, Marcus Widengren, and Anna Wisakanto for discussions and helpful remarks.

Häggström, O. (2016), *Here Be Dragons: Science, Technology and the Future of Humanity* (Oxford University Press).
Häggström, O. (2021), *Tänkande Maskiner: Den Artificiella Intelligensens Genombrott* (Fri Tanke).
Hanson, R. (2009), 'Limit to Growth', *Overcoming Bias*, 21 September, https://www.overcomingbias.com/p/limits-to-growthhtml
Hanson, R. (2019), 'Agency Failure AI Apocalypse?', *Overcoming Bias*, 10 April, https://www.overcomingbias.com/p/agency-failure-ai-apocalypsehtml
Hanson, R. (2023), 'AI Risk, Again', *Overcoming Bias*, 3 March, https://www.overcomingbias.com/p/ai-risk-again
Harris, S., Goldstein, R., and Tegmark, M. (2018), 'What Is and What Matters', *Making Sense Podcast*, 19 March, https://www.samharris.org/podcasts/making-sense-episodes/120-what-is-and-what-matters
Karnofsky, H. (2021a), 'This Can't Go On', *Cold Takes*, 3 August, https://www.cold-takes.com/this-cant-go-on/
Karnofsky, H. (2021b), 'The Most Important Century (in a Nutshell)', *Cold Takes*, 23 September, https://www.cold-takes.com/the-most-important-century-in-a-nutshell/
Kuhn, B. (2014), 'How Many Causes Should You Give To?', https://www.benkuhn.net/how-many-causes/ (accessed August 15, 2022).
Leslie, J. (1998), *The End of the World: The Science and Ethics of Human Extinction* (Routledge).
MacAskill, W. (2019a), 'Are We Living at the Most Influential Time in History?', *Effective Altruism Forum*, 3 September, https://forum.effectivealtruism.org/posts/XXLf6FmWujkxna3E6/are-we-living-at-the-most-influential-time-in-history-1
MacAskill, W. (2019b), Comment 'I Don't Think I Agree with This, Unless . . . ' on MacAskill (2019a), *Effective Altruism Forum*, 13 September, https://forum.effectivealtruism.org/posts/XXLf6FmWujkxna3E6/are-we-living-at-the-most-influential-time-in-history-1
MacAskill, W. (2022a), 'Are We Living at the Hinge of History?', in J. McMahan et al. (eds.) *Ethics and Existence: The Legacy of Derek Parfit* (Oxford University Press), 331–357.
MacAskill, W. (2022b), *What We Owe the Future* (Basic Books).
Mogensen, A. (2022), 'The Hinge of History Hypothesis: Reply to MacAskill', GPI Working Paper No. 9-2022 (Global Priorities Institute, Oxford University).
Murphy, T. (2011), 'Galactic-Scale Energy', *Do the Math*, 12 July, https://dothemath.ucsd.edu/2011/07/galactic-scale-energy/
Ngo. R. (2020), 'AGI Safety From First Principles', https://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view (accessed August 15, 2022).
Ord, T. (2019), Comment 'Hi Will, It is great to see . . . ' on MacAskill (2019a), *Effective Altruism Forum*, 6 September, https://forum.effectivealtruism.org/posts/XXLf6FmWujkxna3E6/are-we-living-at-the-most-influential-time-in-history-1
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Pamlin, D. and Armstrong, S. (2015), *12 Risks That Threaten Human Civilization* (Global Challenges Foundation).
Parfit, D. (1984), *Reasons and Persons* (Clarendon).
Parfit, D. (2011), *On What Matters:* Volume 2 (Oxford University Press).
Schubert, S. (2022), 'Against Cluelessness: Pockets of Predictability', blog post, 18 May, https://stefanschubert.substack.com/p/against-cluelessness-pockets-of-predictability?utm_source=profile&utm_medium=reader2
Shulman, C. (2019), Comment 'I Think This Point Is Even Stronger . . . ' on MacAskill (2019a), *Effective Altruism Forum*, 7 September, https://forum.effectivealtruism.org/posts/XXLf6FmWujkxna3E6/are-we-living-at-the-most-influential-time-in-history-1
Snowden, J. (2019), 'Should We Give to More Than One Charity?', H. Greaves and T. Pummer (eds.), *Effective Altruism: Philosophical Issues* (Oxford University Press), 69–79.
Thomas, T. (2021), 'Doomsday and Objective Chance', GPI Working Paper No. 8-2021 (Global Priorities Institute, Oxford University).
Trammell, P. (2021), 'Dynamic Public Good Provision under Time Preference Heterogeneity: Theory and Applications to Philanthropy', GPI Working Paper No. 9-2021 (Global Priorities Institute, Oxford University).
Williams, D. (1991), *Probability with Martingales* (Cambridge University Press).
Yudkowsky, E. (2022), 'AGI ruin: A List of Lethalities', *LessWrong*, 6 June, https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities

16

How Much Should Governments Pay to Prevent Catastrophes?

Longtermism's Limited Role

Carl Shulman and Elliott Thornley

1 Introduction

It would be very bad if humanity suffered a nuclear war, a deadly pandemic, or an artificial intelligence (AI) disaster. This is for two main reasons. The first is that these catastrophes could kill billions of people. The second is that they could cause human extinction or the permanent collapse of civilization.

Longtermists have argued that humanity should increase its efforts to avert nuclear wars, pandemics, and AI disasters (Beckstead 2013; Bostrom 2013; Ord 2020; Greaves and MacAskill 2021; MacAskill 2022).[1] One prominent longtermist argument for this conclusion appeals to the second reason: these catastrophes could lead to human extinction or the permanent collapse of civilization, and hence prevent an enormous number of potential people from living happy lives in a good future (Beckstead 2013; Bostrom 2013; Ord 2020: 43–49; Greaves and MacAskill 2021; MacAskill 2022: 8–9). These events would then qualify as *existential catastrophes*: catastrophes that destroy humanity's long-term potential (Ord 2020: 37).

Although this longtermist argument has been compelling to many, it has at least two limitations: limitations that are especially serious if the intended conclusion is that *democratic governments* should increase their efforts to prevent catastrophes. First, the argument relies on a premise that many people reject: that it would be an overwhelming moral loss if future generations never exist. Second, the argument overshoots. Given other plausible claims, building policy on this premise would not only lead governments to increase their efforts to prevent catastrophes. It would also lead them to impose extreme costs on the present generation for the sake of miniscule reductions in the risk of existential catastrophe. Since most people's concern for the existence of future generations is limited, this policy would be democratically unacceptable, and so governments cannot use the longtermist argument to guide their catastrophe policy.

In this chapter, we offer a standard cost-benefit analysis argument for reducing the risk of catastrophe. We show that, given plausible estimates of catastrophic risk and the costs of reducing it, many interventions available to governments pass a cost-benefit analysis test.

[1] By 'longtermists', we mean people particularly concerned with ensuring that humanity's long-term future goes well.

Carl Shulman and Elliott Thornley, *How Much Should Governments Pay to Prevent Catastrophes?* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press.
© Carl Shulman and Elliott Thornley 2025. DOI: 10.1093/9780191979972.003.0016

Therefore, the case for averting catastrophe does not depend on longtermism. In fact, we argue, governments should do much more to reduce catastrophic risk even if future generations do not matter at all. The first reason that a catastrophe would be bad—billions of people might die—by itself warrants much more action than the status quo. This argument from present people's interests avoids both limitations of the longtermist argument: it assumes only that the present generation matters, and it does not overshoot. Nevertheless, like the longtermist argument, it implies that governments should do much more to reduce catastrophic risk.

We then argue that getting governments to adopt a catastrophe policy based on cost-benefit analysis should be the goal of longtermists in the political sphere. This goal is achievable, because cost-benefit analysis (CBA) is already a standard tool for government decision-making and because moving to a CBA-driven catastrophe policy would benefit the present generation. Adopting a CBA-driven policy would also reduce the risk of existential catastrophe by almost as much as adopting a *strong longtermist policy* founded on the premise that it would be an overwhelming moral loss if future generations never exist.

We then propose that the longtermist worldview can play a supplementary role in government catastrophe policy. Longtermists can make the case for their view, and thereby increase present people's willingness to pay for *pure longtermist goods*: goods that do not much benefit the present generation but improve humanity's long-term prospects. These pure longtermist goods include especially refuges designed to help civilization recover from future catastrophes. When present people are willing to pay for such things, governments should fund them. This spending would have modest costs for those alive today and great expected benefits for the long-run future.

We end by arguing that longtermists should commit to acting in accordance with a CBA-driven catastrophe policy in the political sphere. This commitment would help bring about an outcome that is much better than the status quo, for the present generation and long-term future alike.

2 The risk of catastrophe

As noted above, we are going to use standard cost-benefit analysis to argue for increased government spending on preventing catastrophes.[2] We focus on the U.S. government, but our points apply to other countries as well (with modifications that will become clear below). We also focus on the risk of *global catastrophes*, which we define as events that kill at least 5 billion people. Many events could constitute a global catastrophe in the coming years, but we concentrate on three in particular: nuclear wars, pandemics, and AI disasters. Reducing the risk of these catastrophes is particularly cost-effective.

The first thing to establish is that the risk is significant. That presents a difficulty. There has never yet been a global catastrophe by our definition, so we cannot base our estimates of the risk on long-run frequencies. But this difficulty is surmountable because we can use other considerations to guide our estimates. These include near-misses (like the Cuban Missile Crisis), statistical models (like power-law extrapolations), and empirical trends

[2] Posner (2004) is one precedent in the literature. In another respect we echo Baum (2015), who argues that we need not appeal to far-future benefits to motivate further efforts to prevent catastrophes.

(like advances in AI). We do not have the space to assess all the relevant considerations in detail, so we mainly rely on previously published estimates of the risks. These estimates should be our point of departure, pending further investigation. Note also that these estimates need not be perfectly accurate for our conclusions to go through. It often suffices that the risks exceed some low value.

Let us begin with the risk of nuclear war. Toby Ord estimates that the existential risk from nuclear war over the next 100 years is about 1 in 1,000 (2020: 167). Note, however, that 'existential risk' refers to the risk of an *existential catastrophe*: a catastrophe that destroys humanity's long-term potential. This is a high bar. It means that any catastrophe from which humanity ever recovers (even if that recovery takes many millennia) does not count as an existential catastrophe. Nuclear wars can be enormously destructive without being likely to pose an existential catastrophe, so Ord's estimate of the risk of 'full-scale' nuclear war is much higher, at about 5% over the next 100 years (Wiblin and Ord 2020). This figure is roughly aligned with our own views (around 3%) and with other published estimates of nuclear risk. At the time of writing, the forecasting community Metaculus puts the risk of thermonuclear war before 2070 at 11% (Metaculus 2025c).[3] Luisa Rodriguez's (2019b) aggregation of expert and superforecaster estimates has the risk of nuclear war between the U.S. and Russia at 0.38% per year, while Martin E. Hellman (2008: 21) estimates that the annual risk of nuclear war between the U.S. and Russia stemming from a Cuban-Missile-Crisis-type scenario is 0.02–0.5%.

We recognize that each of these estimates involves difficult judgment calls. Nevertheless, we think it would be reckless to suppose that the true risk of nuclear war this century is less than 1%. Here are assorted reasons for caution. Nuclear weapons have been a threat for just a single human lifetime, and in those years we have already racked up an eye-opening number of close calls. The Cuban Missile Crisis is the most famous example, but we also have declassified accounts of many accidents and false alarms (see, for example, Ord 2020: Appendix C). And although nuclear conflict would likely be devastating for all sides involved, leaders often have selfish incentives for brinkmanship and may behave irrationally under pressure. Looking ahead, future technological developments may upset the delicate balance of deterrence. And we cannot presume that a nuclear war would harm only its direct targets. Research has suggested that the smoke from smoldering cities would take years to dissipate, during which time global temperatures and rainfall would drop low enough to kill most crops.[4] That leads Rodriguez (2019a) to estimate that a U.S.–Russia nuclear exchange would cause a famine that kills 5.5 billion people in expectation. One of us (Shulman) estimates a lower risk of this kind of *nuclear winter*, a lower average number of warheads deployed in a U.S.–Russia nuclear exchange, and a higher likelihood that emergency measures succeed in reducing mass starvation, but we still put expected casualties in the billions.

Pandemics caused by pathogens that have been engineered in a laboratory are another major concern. Ord (2020: 167) estimates that the existential risk over the next century from these engineered pandemics is around 3%. And as with nuclear war, engineered

[3] A war counts as thermonuclear if and only if three countries each detonate at least 10 nuclear devices of at least 10 kiloton yield outside of their own territory or two countries each detonate at least 50 nuclear devices of at least 10 kiloton yield outside of their own territory.

[4] See, for example, Robock, Oman, and Stenchikov 2007; Mills et al. 2014; Coupe et al. 2019; Xia et al. 2022. Some doubt that nuclear war would have such severe atmospheric effects (Seitz 2011; Reisner et al. 2018).

pandemics could be extremely destructive without constituting an existential catastrophe, so Ord's estimate of the risk of global catastrophe arising from engineered pandemics would be adjusted upward from this 3% figure. At the time of writing, Metaculus suggests that there is a 7.6% probability that an engineered pathogen causes the human population to drop by at least 10% in a period of 5 years or less by 2100.[5] In a 2008 survey of participants at a conference on global catastrophes, the median respondent estimated a 10% chance that an engineered pandemic kills at least 1 billion people and a 2% chance that an engineered pandemic causes human extinction before 2100 (Sandberg and Bostrom 2008).

These estimates are based on a multitude of factors, of which we note a small selection. Diseases can be very contagious and very deadly.[6] There is no strong reason to suppose that engineered diseases could not be both. Scientists continue to conduct research in which pathogens are modified to enhance their transmissibility, lethality, and resistance to treatment (Millett and Snyder-Beattie 2017: 374; Ord 2020: 128–129). We also have numerous reports of lab leaks: cases in which pathogens have been accidentally released from biological research facilities and allowed to infect human populations (Ord 2020: 130–131). Many countries ran bioweapons programs during the twentieth century, and bioweapons were used in both World Wars (Millett and Snyder-Beattie 2017: 374). Terrorist groups like the Aum Shinrikyo cult have tried to use biological agents to cause mass casualties (Millett and Snyder-Beattie 2017: 374). Their efforts were hampered by a lack of technology and expertise, but humanity's collective capacity for bioterror has grown considerably since then. A significant number of people now have the ability to cause a biological catastrophe, and this number looks set to rise further in the coming years (Ord 2020: 133–134).

Ord (2020: 167) puts the existential risk from artificial general intelligence (AGI) at 10% over the next century. This figure is the product of a 50% chance of human-level AGI by 2120 and a 20% risk of existential catastrophe, conditional on AGI by 2120 (Ord 2020: 168–169). Meanwhile, Joseph Carlsmith (2021: 49) estimates a 65% probability that by 2070 it will be possible and financially feasible to build AI systems capable of planning, strategizing, and outperforming humans in important domains. He puts the (unconditional) existential risk from these AI systems at greater than 10% before 2070 (2021: 47). The aggregate forecast in a recent survey of machine learning researchers is a 50% chance of high-level machine intelligence by 2059 (Stein-Perlman, Weinstein-Raun, and Grace 2022).[7] The median respondent in that survey estimated a 5% probability that AI causes human extinction or humanity's permanent and severe disempowerment (Stein-Perlman et al. 2022). Our own

[5] Metaculus forecasters estimate that there is a 33% probability that the human population drops by at least 10% in a period of 5 years or less by 2100 (Metaculus 2025a), and a 23% probability conditional on this drop occurring that it is caused by an engineered pathogen (Metaculus 2025b). Multiplying these figures gets us 7.6%. This calculation ignores some minor technicalities to do with the possibility that there will be more than one qualifying drop in population.

[6] COVID-19 spread to almost every community, as did the 1918 Flu. Engineered pandemics could be even harder to suppress. Rabies and septicemic plague kill almost 100% of their victims in the absence of treatment (Millett and Snyder-Beattie 2017: 374).

[7] The survey defines 'high-level machine intelligence' as machine intelligence that can accomplish every task better and more cheaply than human workers.

Admittedly, we have some reason to suspect these estimates. As Cotra (2020: 40–41) notes, machine learning researchers' responses in a previous survey (Grace et al. 2018) were implausibly sensitive to minor reframings of questions.

In any case, recent progress in AI has exceeded almost all expectations. On two out of four benchmarks, state-of-the-art performance in June 2022 was outside the 90% credible interval of an aggregate of forecasters' predictions made in August 2021 (Steinhardt 2022).

estimates are closer to Carlsmith and the survey respondents on timelines and closer to Ord on existential risk.

These estimates are the most speculative: nuclear weapons and engineered pathogens already exist in the world, while human-level AGI is yet to come. We cannot make a full case for the risk of AI catastrophe in this chapter, but here is a sketch. AI capabilities are growing quickly, powered partly by rapid algorithmic improvements and especially by increasing computing budgets. Before 2010, compute spent on training AI models grew in line with Moore's law, but in the recent deep learning boom it has increased much faster, with an average doubling time of 6 months over that period (Sevilla et al. 2022). Bigger models and longer training runs have led to remarkable progress in domains like computer vision, language, protein modeling, and games. The next 20 years are likely to see the first AI systems close to the computational scale of the human brain, as hardware improves and spending on training runs continues to increase from millions of dollars today to many billions of dollars (Cotra 2020: 1–9; 2022). Extrapolating past trends suggests that these AI systems may also have capabilities matching the human brain across a wide range of domains.

AI developers train their systems using a reward function (or loss function) which assigns values to the system's outputs, along with an algorithm that modifies the system to perform better according to the reward function. But encoding human intentions in a reward function has proved extremely difficult, as is made clear by the many recorded instances of AI systems achieving high reward by behaving in ways unintended by their designers (Krakovna 2018; DeepMind 2020). These include systems pausing Tetris forever to avoid losing (Murphy 2013), using camera trickery to deceive human evaluators into believing that a robot hand is completing a task (OpenAI 2017; DeepMind 2020), and behaving differently under observation to avoid penalties for reproduction (Lehman et al. 2020: 282; Muehlhauser 2021). We also have documented cases of AIs adopting goals that produce high reward in training but differ in important ways from the goals intended by their designers (Langosco et al. 2022; Shah et al. 2022). One example comes in the form of a model trained to win a video game by reaching a coin at the right of the stage. The model retained its ability to navigate the environment when the coin was moved, but it became clear that the model's real goal was to go as far to the right as possible, rather than to reach the coin (Langosco et al. 2022: 4). So far, these issues of *reward hacking* and *goal misgeneralization* have been of little consequence, because we have been able to shut down misbehaving systems or alter their reward functions. But that looks set to change as AI systems come to understand and act in the wider world: a powerful AGI could learn that allowing itself to be turned off or modified is a poor way of achieving its goal (Soares et al. 2015; Thornley 2024). And given any of a wide variety of goals, this kind of AGI would have reason to perform well in training and conceal its real goal until AGI systems are collectively powerful enough to seize control of their reward processes (or otherwise pursue their goals) and defeat any human response (Carlsmith 2023; Ngo et al. 2024).

That is one way in which misaligned AGI could be disastrous for humanity. Guarding against this outcome likely requires much more work on robustly aligning AI with human intentions, along with the cautious deployment of advanced AI to enable proper safety engineering and testing. Unfortunately, economic and geopolitical incentives may lead to much less care than is required. Competing companies and nations may cut corners and expose humanity to serious risks in a race to build AGI (Armstrong, Bostrom, and Shulman 2016). The risk is exacerbated by the *winner's curse* dynamic at play: all else equal, it is the actors

who most underestimate the dangers of deployment that are most likely to do so (Bostrom, Douglas, and Sandberg 2016).

Assuming independence and combining Ord's risk-estimates of 10% for AI, 3% for engineered pandemics, and 5% for nuclear war gives us at least a 17% risk of global catastrophe from these sources over the next 100 years.[8] If we assume that the risk per decade is constant, the risk over the next decade is about 1.85%.[9] If we assume also that every person's risk of dying in this kind of catastrophe is equal, then (conditional on not dying in other ways) each U.S. citizen's risk of dying in this kind of catastrophe in the next decade is at least $5/9 \times 1.85\% \approx 1.03\%$ (since, by our definition, a global catastrophe would kill at least 5 billion people, and the world population is projected to remain under 9 billion until 2035). According to projections of the U.S. population pyramid, 6.45% of U.S. citizens alive today will die in other ways over the course of the next decade.[10] That suggests that U.S. citizens alive today have on average about a 1% risk of being killed in a nuclear war, engineered pandemic, or AI disaster in the next decade. That is about 10 times their risk of being killed in a car accident.[11]

3 Interventions to reduce the risk

There is good reason to think that the risk of global catastrophe in the coming years is significant. Based on Ord's estimates, we suggest that U.S. citizens' risk of dying in a nuclear war, pandemic, or AI disaster in the next decade is on average about 1%. We now survey some ways of reducing this risk.

The Biden administration's 2023 Budget lists many ways of reducing the risk of biological catastrophes (The White House 2022c; U.S. Office of Management and Budget 2022). These include developing advanced personal protective equipment, along with prototype vaccines for the viral families most likely to cause pandemics.[12] The U.S. government can also enhance laboratory biosafety and biosecurity, by improving training procedures, risk

[8] Here we assume that a full-scale nuclear war would kill at least 5 billion people and hence qualify as a global catastrophe (Rodriguez 2019a; Xia et al. 2022).

The risk is not $10\% + 3\% + 5\% = 18\%$, because each of Ord's risk-estimates is conditional on humanity not suffering an existential catastrophe from another source in the next 100 years (as is made clear by Ord 2020: 173–174). If we assume statistical independence between risks, the probability that there is no global catastrophe from AI, engineered pandemics, or nuclear war in the next 100 years is at most $(1-0.1)\times(1-0.03)\times(1-0.05) \approx 83\%$. The probability that there is some such global catastrophe is then at least 17%. There might well be some positive correlation between risks (Ord 2020 : 173–175), but plausible degrees of correlation will not significantly reduce total risk.

Note that the 17% figure does not incorporate the upward adjustment for the (significant, in our view) likelihood that an engineered pandemic constitutes a global catastrophe but not an existential catastrophe.

[9] If the risk over the next century is 17% and the risk per decade is constant, then the risk per decade is x such that $1-(1-x)^{10} = 17\%$. That gives us $x \approx 1.85\%$.

There are reasons to doubt that the risk this decade is as high as the risk in future decades. One might think that 'crunch time' for AI and pandemic risk is more than a decade off. One might also think that most nuclear risk comes from scenarios in which future technological developments cast doubt on nations' second-strike capability, thereby incentivizing first-strikes. These factors are at least partly counterbalanced by the likelihood that we will be better prepared for risks in future decades.

[10] The projected number of Americans at least 10 years old in 2035 is 6.45% smaller than the number of Americans in 2025 (PopulationPyramid 2024).

[11] Our World in Data (2019) records a mean of approximately 41,000 road injury deaths per year in the United States over the past decade.

[12] This Budget includes many of the recommendations from the Apollo Program for Biodefense and Athena Agenda (Bipartisan Commission on Biodefense 2021; 2022).

assessments, and equipment (Bipartisan Commission on Biodefense 2021: 24). Another priority is improving our capacities for microbial forensics (including our abilities to detect engineered pathogens), so that we can better identify and deter potential bad actors (Bipartisan Commission on Biodefense 2021: 24–25). Relatedly, the U.S. government can strengthen the Biological Weapons Convention by increasing the budget and staff of the body responsible for its implementation, and by working to grant them the power to investigate suspected breaches (Ord 2020: 279–280). The Nuclear Threat Initiative recommends establishing a global entity focused on preventing catastrophes from biotechnology, amongst other things (Nuclear Threat Initiative 2020a: 3). Another key priority is developing pathogen-agnostic detection technologies. One such candidate technology is a Nucleic Acid Observatory, which would monitor waterways and wastewater for changing frequencies of biological agents, allowing for the early detection of potential biothreats (The Nucleic Acid Observatory Consortium 2021).

The U.S. government can also reduce the risk of nuclear war this decade. Ord (2020: 278) recommends restarting the Intermediate-Range Nuclear Forces Treaty, taking U.S. intercontinental ballistic missiles off of hair-trigger alert ('Launch on Warning'), and increasing the capacity of the International Atomic Energy Agency to verify that nations are complying with safeguards agreements. Other recommendations come from the Centre for Long-Term Resilience's *Future Proof* report (2021). They are directed towards the U.K. government but apply to the U.S. as well. The recommendations include committing not to incorporate AI systems into nuclear command, control, and communications (NC3) and lobbying to establish this norm internationally.[13] Another is committing to avoid cyber operations that target the NC3 of Non-Proliferation Treaty signatories and establishing a multilateral agreement to this effect. The Nuclear Threat Initiative (2020b) offers many recommendations to the Biden administration for reducing nuclear risk, some of which have already been taken up.[14] Others include working to bring the Comprehensive Nuclear-Test-Ban Treaty into force, re-establishing the Joint Comprehensive Plan of Action's limits on Iran's nuclear activity, and increasing U.S. diplomatic efforts with Russia and China (Nuclear Threat Initiative 2020b).[15]

To reduce the risks from AI, the U.S. government can fund research in AI safety.[16] This should include alignment research focused on reducing the risk of catastrophic AI takeover by ensuring that even very powerful AI systems do what we intend, as well as interpretability research to help us understand neural networks' behavior and better supervise their training (Amodei et al. 2016; Hendrycks et al. 2022). The U.S. government can also fund research and work in AI governance, focused on devising norms, policies, and institutions to ensure that the development of AI is beneficial for humanity (Dafoe 2018).

[13] Avin and Amadae (2019) survey ways in which AI may exacerbate nuclear risk and offer policy recommendations, including the recommendation not to incorporate AI into NC3. The U.S. National Security Commission on Artificial Intelligence (2021: 98) makes a similar recommendation.

[14] Those taken up already include extending New START (Strategic Arms Reduction Treaty) and issuing a joint declaration with the other members of the P5—China, France, Russia, and the U.K.—that a 'nuclear war cannot be won and must never be fought' (The White House 2022b).

[15] It is worth noting that the dynamics of nuclear risk are complex, and that experts disagree about the likely effects of these interventions. What can be broadly agreed is that nuclear risk should receive more investigation and funding.

[16] The National Science Foundation's $20 million in grants for AI safety research is a promising step in this direction (National Science Foundation 2023).

4 Cost-benefit analysis of catastrophe-preventing interventions

We project that funding this suite of interventions for the next decade would cost less than US$400 billion.[17] We also expect this suite of interventions to reduce the risk of global catastrophe over the next decade by at least 0.1pp (percentage points). A full defense of this claim would require more detail than we can fit in this chapter, but here is one way to illustrate the claim's plausibility. Imagine an enormous set of worlds like our world in 2025. Each world in this set is different with respect to the features of our world about which we are uncertain, and worlds with a certain feature occur in the set in proportion to our best evidence about the presence of that feature in our world. If, for example, the best appraisal of our available evidence suggests that there is a 55% probability that the next U.S. President will be a Democrat, then 55% of the worlds in our set have a Democrat as the next President. We claim that *in at least 1 in 1,000 of these worlds* the interventions we recommend above would prevent a global catastrophe this decade. That is a low bar, and it seems plausible to us that the interventions above meet it. Our question now is: given this profile of costs and benefits, do these interventions pass a standard cost-benefit analysis test?

To assess interventions expected to save lives, cost-benefit analysis begins by *valuing mortality risk reductions*: putting a monetary value on reducing citizens' risk of death (Kniesner and Viscusi 2019). To do that, we first determine how much a representative sample of citizens are willing to pay to reduce their risk of dying this year by a given increment (often around 0.01pp, or 1 in 10,000). One method is to ask them, giving us their stated preferences. Another method is to observe people's behavior, particularly their choices about what to buy and what jobs to take, giving us their revealed preferences.[18]

U.S. government agencies use methods like these to estimate how much U.S. citizens are willing to pay to reduce their risk of death.[19] This figure is then used to calculate the *value of a statistical life* (VSL): the value of saving one life in expectation via small reductions in mortality risks for many people. The primary VSL figure used by the U.S. Department of Transportation for 2021 is $11.8 million, with a range to account for various kinds of uncertainty spanning from about $7 million to $16.5 million (U.S. Department of Transportation 2021a; 2021b).[20] These figures are used in the cost-benefit analyses of policies expected to save lives. Costs and benefits occurring in the future are discounted at a constant annual rate. The Environmental Protection Agency (EPA) uses annual discount rates of 2% and 3%; the Office of Information and Regulatory Affairs (OIRA) instructs agencies to conduct

[17] The Biden administration's 2023 Budget requests $88.2 billion over five years (The White House 2022c; U.S. Office of Management and Budget 2022). We can suppose that another five years of funding would require that much again. A Nucleic Acid Observatory covering the U.S. is estimated to cost $18.4 billion to establish and $10.4 billion per year to run (The Nucleic Acid Observatory Consortium 2021: 18). Ord (2020: 202–203) recommends increasing the budget of the Biological Weapons Convention to $80 million per year. Our listed interventions to reduce nuclear risk are unlikely to cost more than $10 billion for the decade. AI safety and governance might cost up to $10 billion as well. The total cost of these interventions for the decade would then be $319.6 billion.

[18] We can observe how much people pay for products that reduce their risk of death, like bike helmets, smoke alarms, and airbags. We can also observe how much more people are paid to do risky work, like service nuclear reactors and fly new planes (Kniesner and Viscusi 2019).

[19] U.S. agencies rely mainly on hedonic wage studies, which measure the wage-premium for risky jobs. European agencies tend to rely on stated preference methods (Kniesner and Viscusi 2019: 10).

[20] Updating for inflation and growth in real incomes, the U.S. Environmental Protection Agency's central estimate for 2021 is approximately $12.2 million. The U.S. Department of Health and Human Services' 2021 figure is about $12.1 million (Kniesner and Viscusi 2019).

analyses using annual discount rates of 3% and 7% (Graham 2008: 504). The rationale is opportunity costs and people's rate of pure time preference (Graham 2008: 504).

Now for the application to the risk of global catastrophe (otherwise known as *global catastrophic risk*, or *GCR*). We defined a global catastrophe above as an event that kills at least 5 billion people, and we assumed that each person's risk of dying in a global catastrophe is equal. So, given a world population of less than 9 billion and conditional on a global catastrophe occurring, each American's risk of dying in that catastrophe is at least 5/9. Reducing GCR this decade by 0.1pp then reduces each American's risk of death this decade by at least 0.055pp. Multiplying that figure by the U.S. population of 330 million, we get the result that reducing GCR this decade by 0.1pp saves at least 181,500 American lives in expectation. If that GCR-reduction were to occur this year, it would be worth at least $1.27 trillion on the Department of Transportation's lowest VSL figure of $7 million. But since the GCR-reduction would occur over the course of a decade, cost-benefit analysis requires that we discount. If we use OIRA's highest annual discount rate of 7% and suppose (conservatively) that all the costs of our interventions are paid up front while the GCR-reduction comes only at the end of the decade, we get the result that reducing GCR this decade by 0.1pp is worth at least $1.27 trillion $/1.07^{10}$ = $646 billion. So, at a cost of $400 billion, these interventions comfortably pass a standard cost-benefit analysis test.[21] That in turn suggests that the U.S. government should fund these interventions. Doing so would save American lives more cost-effectively than many other forms of government spending on life-saving, such as transportation and environmental regulations.

In fact, we can make a stronger argument. Using a projected U.S. population pyramid and some life-expectancy statistics, we can calculate that approximately 80% of the American life-years saved by preventing a global catastrophe in 2035 would accrue to Americans alive today in 2025 (Thornley 2025). Eighty percent of $646 billion is approximately $517 billion. That means that funding this suite of GCR-reducing interventions is well worth it, even considering only the benefits to Americans alive today.

And recall that the above figures assume a 0.1pp reduction in GCR as a result of implementing the whole suite of interventions. In our judgment, a 0.5pp reduction in GCR is a better estimate, in which case the benefit-cost ratio of the suite is over 5. Making our other assumptions more reasonable results in even more favorable benefit-cost ratios. Using the Department of Transportation's primary VSL figure of $11.8 million and an annual discount rate of 3%, the benefit-cost ratio of the suite comes out at over 20.[22] The most

[21] Researchers and analysts in the U.S. frequently cite a $50,000-per-quality-adjusted-life-year (QALY) threshold for funding medical interventions, but this figure lacks any particular normative significance and has not been updated to account for inflation and real growth in incomes since it first came to prominence in the mid-1990s (Neumann, Cohen, and Weinstein 2014). The £20,000–30,000-per-QALY range recommended by the U.K.'s National Institute for Health and Care Excellence suffers from similar defects (Claxton et al. 2016). More principled estimates put a higher value on years of life (Hirth et al. 2000; Aldy and Viscusi 2008; Favaloro and Berger 2021). In any case, simply updating the $50,000-per-QALY threshold to account for inflation and growth since 1995 would imply a value of more than $100,000-per-QALY. At $100,000-per-QALY, the value of reducing GCR a decade from now by 0.1pp is at least $0.001 \times \$100,000 \times 14,583,317,092 \times (5/9) \times (1/1.07^{10}) \approx \$412$ billion. (14,583,317,092 is the expected number of American life-years saved by preventing a global catastrophe in 2035, based on a projected U.S. population pyramid (PopulationPyramid 2024) and life-expectancy statistics (U.S. Social Security Administration 2022). See Thornley (2025).) That figure justifies the suite of interventions we recommend below. We believe that many interventions are also justified on the more demanding $50,000-per-QALY figure.

[22] $0.005 \times \$11,800,000 \times 330,000,000 \times (5/9) \times (1/1.03^{10}) \approx \$8.05$ trillion, which is over 20 times the cost of $400 billion.

cost-effective interventions within the suite will have benefit-cost ratios that are more favorable still.

It is also worth noting some important ways in which our calculations up to this point underrate the value of GCR-reducing interventions. First, we have appealed only to these interventions' GCR-reducing benefits: the benefits of shifting probability mass away from outcomes in which at least 5 billion people die and towards outcomes in which very few people die. But these interventions would also decrease the risk of smaller catastrophes, in which less than 5 billion people die.[23] Second, the value of preventing deaths from catastrophe is plausibly higher than the value of preventing traffic deaths. The EPA (2010: 20–26) and U.K. Treasury (2003: 62) have each recommended that a higher VSL be used for cancer risks than for accidental risks, to reflect the fact that dying from cancer tends to be more unpleasant than dying in an accident (Kniesner and Viscusi 2019: 16). We suggest that the same point applies to death by nuclear winter and engineered pandemic.

Here is another benefit of our listed GCR-reducing interventions. They do not just reduce U.S. citizens' risk of death. They also reduce the risk of death for citizens of other nations. That is additional reason to fund these interventions.[24] It also suggests that the U.S. government could persuade other nations to share the costs of GCR-reducing interventions, in which case funding these interventions becomes an even more cost-effective way of saving U.S. lives. Cooperation between nations can also make it worthwhile for the U.S. and the world as a whole to spend more on reducing GCR. Suppose, for example, that there is some intervention that would cost $1 trillion and would reduce GCR by 0.1pp over the next decade. That is too expensive for the U.S. alone (at least based on our conservative calculations), but it would be worth funding for a coalition of nations that agreed to split the cost.

5 Longtermists should advocate for a CBA-driven catastrophe policy

The U.S. is seriously underspending on preventing catastrophes. This conclusion follows from standard cost-benefit analysis. We need not be longtermists to believe that

[23] This is especially so in the case of pandemics, and in fact the pandemic-preventing interventions that we list are justified even considering only their effects on the risk of pandemics about as damaging as COVID-19. The total cost of the COVID-19 pandemic for the U.S. has been estimated at $16 trillion (Cutler and Summers 2020), which suggests that it is worth the U.S. spending up to $32 billion per year to decrease the annual risk of such pandemics by 0.2pp (and Cutler and Summers' estimate is based on an October 2020 projection of 625,000 deaths. At the time of writing, Our World in Data (2025) has total confirmed U.S. COVID-19 deaths at over 1.2 million). Our listed pandemic-preventing interventions are projected to cost less than $32 billion per year, and they would plausibly reduce annual risk by more than 0.2pp. After all, the observed frequency of pandemics as bad as COVID-19 is about one per century, suggesting an annual risk of 1% per year. A 0.2pp decrease then means a 20% decrease in baseline risk, which seems easily achievable via the interventions that we recommend. And since our listed pandemic-preventing interventions can be justified in this way, the case for funding them does not depend on difficult forecasts of the likelihood of unprecedented events, like a pandemic constituting a global catastrophe. Instead, we can appeal to the observed frequency of pandemics about as damaging as COVID-19.

[24] There is a case for including benefits to non-U.S. citizens in cost-benefit analyses of GCR-reducing interventions. After all, saving the lives of non-U.S. citizens is morally important. And the Biden administration already includes costs to non-U.S. citizens in its social cost of carbon (SCC): its estimate of the harm caused by carbon dioxide emissions (The White House 2022a). The SCC is a key input to the U.S. government's climate policy, and counting costs to non-U.S. citizens in the SCC changes the cost-benefit balance of important decisions like regulating power plant emissions, setting standards for vehicle fuel efficiency, and signing on to international climate agreements.

the U.S. government should do much more to reduce the risk of nuclear wars, pandemics, and AI disasters. In fact, even entirely self-interested Americans have reason to hope that the U.S. government increases its efforts to avert catastrophes. The interventions that we recommend above are well worth it, even considering only the benefits to Americans alive today. Counting the benefits to citizens of other nations and the next generation makes these interventions even more attractive. So, Americans should hope that the U.S. government adopts something like a *CBA-driven catastrophe policy*: a policy of funding all those GCR-reducing interventions that pass a cost-benefit analysis test.

One might think that longtermists should be more ambitious: that rather than push for a CBA-driven catastrophe policy, longtermists should urge governments to adopt a *strong longtermist policy*. By a 'strong longtermist policy', we mean a policy founded on the premise that it would be an overwhelming moral loss if future generations never exist.[25] However, we argue that this is not the case: longtermists should advocate for a CBA-driven catastrophe policy rather than a strong longtermist policy. That is because (i) unlike a strong longtermist policy, a CBA-driven policy would be democratically acceptable and feasible to implement, and (ii) a CBA-driven policy would reduce existential risk by almost as much as a strong longtermist policy.[26]

Let us begin with democratic acceptability. As noted above, a strong longtermist policy would in principle place extreme burdens on the present generation for the sake of even miniscule reductions in existential risk. Here is a rough sketch of why. If the non-existence of future generations would be an overwhelming moral loss, then an existential catastrophe (like human extinction or the permanent collapse of civilization) would be extremely bad. That in turn makes it worth reducing the risk of existential catastrophe even if doing so is exceedingly costly for the present generation.[27]

We now argue that a strong longtermist policy would place serious burdens on the present generation not only in principle but also in practice. There are suites of existential-risk-reducing interventions that governments could implement only at extreme cost to those alive today. For example, governments could slow down the development of existential-risk-increasing technologies (even those that pose only very small risks) by paying researchers large salaries to do other things. Governments could also build extensive, self-sustaining colonies (in remote locations or perhaps far underground) in which residents are permanently cut off from the rest of the world and trained to rebuild civilization in the event of a

[25] Describing this policy as 'longtermist' is simplifying slightly. Some longtermists prioritize preventing future suffering over increasing the probability that future generations exist (see, for example, Vinding 2020).

[26] Here is a related recommendation: longtermists should assess interventions' cost-effectiveness using standard cost-benefit analysis when proposing those interventions to governments. They should not assess cost-effectiveness using longtermist assumptions and then appeal to cost-effectiveness thresholds from standard cost-benefit analysis to argue for government funding (see, e.g., Matheny 2007: 1340). If governments funded every intervention justified on these grounds, their level of spending on catastrophe-preventing interventions would be unacceptable to a majority of their citizens.

[27] Bostrom (2013: 18–19) makes something like this point, as does Posner (2004: 152–153).
Why think that the non-existence of future generations would be an overwhelming moral loss? The best-known argument goes as follows: the expected future population is enormous (Greaves and MacAskill 2021: 6–9; MacAskill 2022: 1), the lives of future people are good in expectation (MacAskill 2022: 9), and—all else equal—it is better if the future contains more good lives (MacAskill 2022: 8). We should note, however, that longtermism is a big tent and that not all longtermists accept these claims.

catastrophe. The U.S. government could set up a *global* Nucleic Acid Observatory, paying other countries large fees (if need be) to allow the U.S. to monitor their water supplies for emerging pathogens. More generally, governments could heavily subsidize investment, research, and development in ways that incentivize the present generation to increase civilization's resilience and decrease existential risk. A strong longtermist policy would seek to implement these and other interventions quickly, a factor which adds to their expense. These expenses would in turn require increasing taxes on present citizens (particularly consumption taxes), as well as cutting forms of government spending that have little effect on existential risk (like Social Security, many kinds of medical care, and funding for parks, art, culture, and sport). These budget changes would be burdensome for those alive today. Very cautious regulation of technological development would impose burdens too. It might mean that present citizens miss out on technologies that would improve and extend their lives, like consumer goods and cures for diseases.

So, a strong longtermist policy would be *democratically unacceptable*, by which we mean it could not be adopted and maintained by a democratic government. If a government tried to adopt a strong longtermist policy, it would lose the support of most of its citizens. There are clear moral objections against governments implementing democratically unacceptable policies, but even setting those aside, getting governments to adopt a strong longtermist policy is not feasible. Efforts in that direction are very unlikely to succeed.

A CBA-driven catastrophe policy, by contrast, would be democratically acceptable. This kind of policy would not place heavy burdens on the present generation. Since cost-benefit analysis is based in large part on citizens' willingness to pay, policies guided by cost-benefit analysis tend not to ask citizens to pay much more than is in their own interests. And given our current lack of spending on preventing catastrophes, moving from the status quo to a CBA-driven policy is almost certainly good for U.S. citizens alive today. That is one reason to think that getting the U.S. government to adopt a CBA-driven policy is particularly feasible. Another is that cost-benefit analysis is already a standard tool for U.S. regulatory decision-making.[28] Advocating for a CBA-driven policy does not mean asking governments to adopt a radically new decision-procedure. It just means asking them to extend a standard decision-procedure into a domain where it has so far been underused.

Of course, getting governments to adopt a CBA-driven catastrophe policy is not trivial. One barrier is psychological (Wiener 2016). Many of us find it hard to appreciate the likelihood and magnitude of a global catastrophe. Another is that GCR-reduction is a collective action problem for individuals. Although a safer world is in many people's self-interest, *working* for a safer world is in few people's self-interest. Doing so means bearing a large portion of the costs and gaining just a small portion of the benefits.[29] Politicians and regulators

[28] Since the Reagan administration, executive orders have required U.S. agencies to conduct cost-benefit analyses of major regulations (The U.S. National Archives and Records Administration 2012), and to demonstrate that the benefits of the regulation outweigh the costs (U.S. Government 1982). U.S. courts have struck down regulations for being insufficiently sensitive to the results of cost-benefit analyses (Graham 2008: 454, 479; Posner and Sunstein 2017: 1820), citing a clause in the Administrative Procedure Act which requires courts to invalidate regulations that are 'arbitrary [or] capricious' (Scope of Review 2012). The Supreme Court has indicated that agencies may not impose regulations with costs that 'significantly' exceed benefits (Michigan v. EPA 2015). For more, see Graham (2008) and Posner and Sunstein (2017).

[29] In this respect, reducing GCR is akin to mitigating climate change.

likewise lack incentives to advocate for GCR-reducing interventions (as they did with climate interventions in earlier decades). Given widespread ignorance of the risks, calls for such interventions are unlikely to win much public favor.

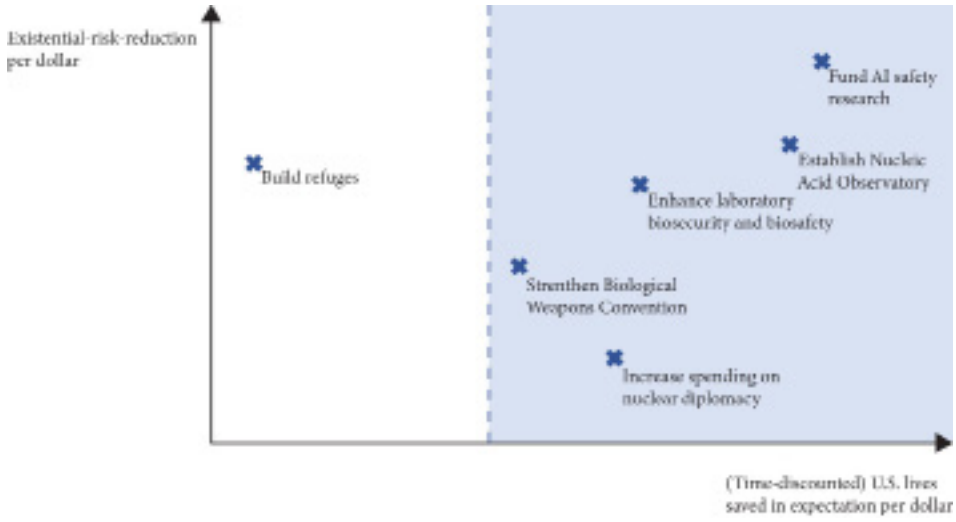
However, these barriers can be overcome. Those willing to bear costs for the sake of others can use their time and money to make salient the prospect of global catastrophe, thereby fostering public support for GCR-reducing interventions and placing them on the policy agenda.[30] Longtermists—who care about the present generation as well as future generations—are well-suited to play this role in pushing governments to adopt a CBA-driven catastrophe policy. If they take up these efforts, they have a good chance of succeeding.

Now for the second point: getting the U.S. government to adopt a CBA-driven catastrophe policy would reduce existential risk by almost as much as getting them to adopt a strong longtermist policy. This is for two reasons. The first is that, at the current margin, the primary goals of a CBA-driven policy and a strong longtermist policy are substantially aligned. The second is that increased spending on preventing catastrophes yields steeply diminishing returns in terms of existential-risk-reduction.

Let us begin with substantial alignment. The primary goal of a CBA-driven catastrophe policy is saving lives in the near term. The primary goal of a strong longtermist policy is reducing existential risk. In the world as it is today, these goals are aligned: many of the best interventions for reducing existential risk are also cost-effective interventions for saving lives in the near term. Take AI, for example. Per Ord (2020: 167) and many other longtermists, the risk from AI makes up a large portion of the total existential risk this century, and this risk could be reduced significantly by work on AI safety and governance. That places this work high on many longtermists' list of priorities. We have argued above that a CBA-driven policy would also fund this work, since it is a cost-effective way of saving lives in the near term. The same goes for pandemics. Interventions to thwart potential pandemics rank highly on the longtermist list of priorities, and these interventions would also be implemented by a CBA-driven policy.

We illustrate the alignment between a CBA-driven policy and a strong longtermist policy using the graph below. The x-axis represents U.S. lives saved (discounted by how far in the future the life is saved) in expectation per dollar. The y-axis represents existential-risk-reduction per dollar. Interventions to the right of the blue line would be funded by a CBA-driven catastrophe policy. The exact position of each intervention is provisional and unimportant, and the graph is not to scale in any case. The important point is that a CBA-driven policy would fund many of the best interventions for reducing existential risk.

[30] There have recently been some promising steps in this direction. In late 2022, the Global Catastrophic Risk Management Act passed committee consideration and was placed on the U.S. Senate's legislative calendar (S.4488 - Global Catastrophic Risk Management Act of 2022). The bill would require the President to establish an interagency committee on global catastrophic risk and to submit to Congress a detailed assessment of the risks, including expert estimates and recommendations for action.


Existential-risk-reduction per dollar
Build refuges
Fund AI safety research
Establish Nucleic Acid Observatory
Enhance laboratory biosecurity and biosafety
Strengthen Biological Weapons Convention
Increase spending on nuclear diplomacy
(Time-discounted) U.S. lives saved in expectation per dollar

That is the key alignment between a CBA-driven policy and a strong longtermist policy. Now for three potentially significant differences. The first is that a strong longtermist policy would fund what we call *pure longtermist goods*: goods that do not much benefit present people but improve humanity's long-term prospects. These pure longtermist goods include refuges to help humanity recover from catastrophes. The second difference is that a strong longtermist policy would spend much more on preventing catastrophes than a CBA-driven policy. In addition to the interventions warranted by a CBA-driven catastrophe policy, a strong longtermist policy would also fund catastrophe-preventing interventions that are too expensive to pass a cost-benefit analysis test. The third difference concerns nuclear risks. The risk of a full-scale nuclear war is significantly higher than the risk of a nuclear war constituting an existential catastrophe (5% versus 0.1% this century, per Ord). In part for this reason, interventions to reduce nuclear risk are cost-effective for saving lives in the near term but not so cost-effective for reducing existential risk.[31] That makes these interventions a relatively lower priority by the lights of a strong longtermist policy than they are by the lights of a CBA-driven policy. Holding fixed the catastrophe-budget warranted by cost-benefit analysis, a strong longtermist policy would likely shift some funding away from nuclear interventions and towards AI and pandemic interventions that fail a cost-benefit analysis test.[32]

Set aside pure longtermist goods for now. We discuss them in the next section. Consider instead the fact that a strong longtermist policy would spend considerably more on preventing catastrophes (especially AI and biological catastrophes) than a CBA-driven policy. We argue that this extra spending would not make such a significant difference

[31] However, we should note that nuclear war is an *existential risk factor* (Ord 2020: 175–180): a factor that increases existential risk. That is because nuclear wars that are not themselves existential catastrophes make humanity more vulnerable to other kinds of existential catastrophe. Since nuclear war is an existential risk factor, preventing nuclear war has effects on total existential risk not limited by nuclear war's direct contribution to existential risk.

[32] We should note, though, that there are other reasons why a strong longtermist policy might prioritize nuclear risk. One is that a nuclear war might negatively affect the characteristics of the societies that shape the future.

to existential risk, because increased spending on preventing catastrophes yields steeply diminishing returns in terms of existential-risk-reduction. That in turn is for two primary reasons. The first is that the most promising existential-risk-reducing interventions—for example, AI safety and governance, a Nucleic Acid Observatory, enhanced biosecurity and biosafety practices—pass a cost-benefit analysis test. Those catastrophe-preventing interventions that fail a cost-benefit analysis test are not nearly as effective in reducing existential risk.

Here is a second reason to expect increased spending to yield steeply diminishing returns in terms of existential-risk-reduction: many interventions *undermine* each other. What we mean here is that many interventions render other interventions less effective, so that the total existential-risk-reduction gained by funding some sets of interventions is less than the sum of the existential-risk-reduction gained by funding each intervention individually. Consider an example. Setting aside a minor complication, we can decompose existential risk from engineered pathogens into two factors: the risk that an engineered pathogen infects more than 1,000 people, and the risk of an existential catastrophe given that an engineered pathogen infects more than 1,000 people.[33] Suppose (for the sake of illustration only) that each risk is 10% this decade, that incentivizing the world's biomedical researchers to do safer research would halve the first risk, and that establishing a Nucleic Acid Observatory (NAO) would halve the second risk. Then in the absence of any interventions, existential risk this decade from engineered pathogens is 1%. Only incentivizing safe research would reduce existential risk by 0.5%. Only establishing an NAO would reduce existential risk by 0.5%. But incentivizing safe research *after* establishing an NAO reduces existential risk by just 0.25%. More generally, the effectiveness of existential-risk-reducing interventions that fail a cost-benefit analysis test would be substantially undermined by all those interventions that pass a cost-benefit analysis test.

At the moment, the world is spending very little on preventing global catastrophes. The U.S. spent approximately $3 billion on biosecurity in 2019 (Watson et al. 2018), and (in spite of the wake-up call provided by COVID-19) funding for preventing future pandemics has not increased much since then.[34] Much of this spending is ill-suited to combatting the most extreme biological threats. Spending on reducing GCR from AI is less than $100 million per year.[35] So, there is a lot of low-hanging fruit for governments to pick: given the current lack of spending, moving to a CBA-driven catastrophe policy would significantly decrease existential risk. Governments could reduce existential risk further by moving to a strong longtermist policy, but this extra reduction would be comparatively small. The same goes for shifting funding away from nuclear risk and towards AI and pandemic risks while holding fixed the level of spending on catastrophe-prevention warranted by cost-benefit analysis. This shift would have just a small effect on existential risk, because the best

[33] The minor complication is that an engineered pathogen could cause an existential catastrophe (the destruction of humanity's long-term potential) *without* infecting more than 1,000 people. Since this outcome is very unlikely, we can safely ignore it here.

[34] The PREVENT Pandemics Act (S.3799 - PREVENT Pandemics Act 2022) includes only about $2 billion in new spending to prevent future pandemics. Biden's Build Back Better Act originally included $2.7 billion of funding for pandemic prevention (Teran 2022), but this funding was cut when the legislation became the Inflation Reduction Act (H.R.5376 - Inflation Reduction Act of 2022).

[35] Ord (2020: 312) estimated that global spending on reducing existential risk from AI in 2020 was between $10 and $50 million per year.

interventions for reducing AI and pandemic risks would already have been funded by a CBA-driven policy.

And, as noted above, international cooperation would make even more catastrophe-preventing interventions cost-effective enough to pass a cost-benefit analysis test. Some of these extra interventions would also have non-trivial effects on existential risk. Consider climate change. Some climate interventions are too expensive to be in any nation's self-interest to fund unilaterally but are worth funding for a coalition of nations that agree to coordinate. Transitioning from fossil fuels to renewable energy sources is one example. Climate change is also an *existential risk factor*: a factor that increases existential risk. Besides posing a small risk of directly causing human extinction or the permanent collapse of civilization, climate change poses a significant indirect risk. It threatens to exacerbate international conflict and drive humanity to pursue risky technological solutions. Extreme climate change would also damage our resilience and make us more vulnerable to other catastrophes. So, in addition to its other benefits, mitigating climate change decreases existential risk. Since more climate interventions pass a cost-benefit analysis test if nations agree to coordinate, this kind of international cooperation would further shrink the gap between existential risk on a CBA-driven catastrophe policy versus a strong longtermist policy.

6 Pure longtermist goods and altruistic willingness to pay

There remains one potentially important difference between a CBA-driven catastrophe policy and a strong longtermist policy: a strong longtermist policy will provide significant funding for what we call *pure longtermist goods*. These we define as goods that do not much benefit the present generation but improve humanity's long-term prospects. They include especially *refuges*: large, well-equipped structures akin to bunkers or shelters, designed to help occupants survive future catastrophes and later rebuild civilization.[36] It might seem like a CBA-driven catastrophe policy would provide no funding for pure longtermist goods, because they are not particularly cost-effective for saving lives in the near term. In the event of a serious catastrophe, refuges would save at most a small portion of the people alive today. But a strong longtermist policy would invest in refuges, because they would significantly reduce existential risk. Even a relatively small group of survivors could get humanity back on track, in which case an existential catastrophe—the permanent destruction of humanity's long-term potential—will have been averted. Since a strong longtermist policy would provide funding for refuges, it might seem as if adopting a strong longtermist policy would reduce existential risk by significantly more than adopting a CBA-driven policy.

However, even this difference between a CBA-driven policy and a strong longtermist policy need not be so great. That is because cost-benefit analysis should incorporate (and is beginning to incorporate) citizens' willingness to pay to uphold their moral commitments: what we will call their *altruistic willingness to pay* (AWTP). Posner and Sunstein (2017) offer arguments to this effect. They note that citizens have various moral commitments—concerning the natural world, non-human animals, citizens of other nations, future generations, etc.—and suffer welfare losses when these commitments are

[36] See Beckstead (2015) and Jebari (2015) for more detail.

compromised (2017: 1829–1830).[37] They argue that the best way to measure these losses is by citizens' willingness to pay to uphold their moral commitments, and that this willingness to pay should be included in cost-benefit calculations of proposed regulations (2017: 1830).[38] Posner and Sunstein also note that there is regulatory and legal precedent for doing so (2017: sec. 3).[39]

And here, we believe, is where longtermism should enter into government catastrophe policy. Longtermists should make the case for their view, and thereby increase citizens' AWTP for pure longtermist goods like refuges.[40] When citizens are willing to pay for these goods, governments should fund them.

Although the uptake of new moral movements is hard to predict (Sunstein 2020), we have reason to be optimistic about this kind of longtermist outreach. A recent survey suggests that many people have moral intuitions that might incline them towards a weak form of longtermism: respondents tended to judge that it's good to create happy people (Caviola et al. 2022: 9). Another survey indicates that simply making the future salient has a marked effect on people's views about human extinction. When prompted to consider long-term consequences, the proportion of people who judged human extinction to be uniquely bad relative to near-extinction rose from 23% to 50% (Schubert, Caviola, and Faber 2019: 3–4). And when respondents were asked to suppose that life in the future would be much better than life today, that number jumped to 77% (Schubert et al. 2019: 4). In the span of about six decades, environmentalism has grown from a fringe movement to a major moral priority of our time. Like longtermism, it has been motivated in large part by a concern for

[37] The welfare loss is most direct on an unrestricted preference-satisfaction theory of welfare: if a person has a moral commitment compromised, they thereby have a preference frustrated and so suffer a welfare loss. But compromised moral commitments also lead to welfare losses on other plausible theories of welfare. These theories will place some weight on positive and negative experiences, and having one's moral commitments compromised is typically a negative experience.

[38] Here are two reasons why one might think that AWTP should be excluded from cost-benefit calculations, along with responses. First, one might think that AWTP for benefits to other people should be excluded (U.S. Environmental Protection Agency 2010: 18–19). Most of us care not only about the benefits that other people receive, but also about the costs that they bear. If benefits but not costs are included, we all pay more for benefits than we would like to, on average. If both benefits and costs are included, they cancel each other out. This point is correct as far as it goes, but it gives us no reason to exclude AWTP for pure longtermist goods from cost-benefit calculations. Future generations will not have to pay for the pure longtermist goods that we fund (U.S. Environmental Protection Agency 2010: 19).

Second, one might think that charities (rather than governments) should assume the responsibility of upholding citizen's moral commitments. This thought is analogous to the thought that private companies (rather than governments) should provide for citizens' needs, and the response is analogous as well: some collective action problems require government action to solve. Citizens may be willing to bear costs for the sake of some moral commitment if and only if it can be ensured that some number of other people are contributing as well (Posner and Sunstein 2017: 1840).

[39] In their cost-benefit analysis of the 'Nondiscrimination on the Basis of Disability in State and Local Government Services' regulation, the U.S. Department of Justice (DOJ) appealed to non-wheelchair-users' willingness to pay to make buildings more accessible for wheelchair users. The DOJ noted that, even if non-wheelchair-users would be willing to pay just pennies on average to provide disabled access, the benefits of the regulation would justify the costs (Nondiscrimination on the Basis of Disability in State and Local Government Services 2010). In another context, the DOJ estimated U.S. AWTP to prevent rape, and noted that the estimated figure justified a regulation designed to reduce the incidence of prison rape (National Standards to Prevent, Detect, and Respond to Prison Rape 2012). And on the legal side, the U.S. Department of the Interior had a damage measure struck down by a court of appeals for failing to incorporate the *existence value* of pristine wilderness: the value that people derive from just knowing that such places exist, independently of whether they expect to visit them (*Ohio v. U.S. Dept. Of the Interior* 1989). Based on this case, Sunstein and Posner (2017: 1858–1860) suggest that excluding AWTP from cost-benefit analyses may suffice to render regulations 'arbitrary [and] capricious', in which case courts are required by the Administrative Procedure Act to invalidate them (Scope of Review 2012).

[40] Baum (2015: 93) makes a point along these lines: longtermists can use the inspirational power of the far future to motivate efforts to ensure it goes well.

future generations. Longtermist arguments have already been compelling to many people, and these factors suggest that they could be compelling to many more.

Even a small AWTP for pure longtermist goods could have a significant effect on existential risk. If U.S. citizens are willing to contribute just $5 per year on average, then a CBA-driven policy that incorporates AWTP warrants spending up to $1.65 billion per year on pure longtermist goods: enough to build extensive refuges. Of course, even in a scenario in which every U.S. citizen hears the longtermist arguments, a CBA-driven policy will provide less funding for pure longtermist goods than a strong longtermist policy. But, as with catastrophe-preventing interventions, it seems likely that marginal existential-risk-reduction diminishes steeply as spending on pure longtermist goods increases: so steeply that moving to the level of spending on pure longtermist goods warranted by citizens' AWTP would reduce existential risk by almost as much as moving to the level of spending warranted by a strong longtermist policy. This is especially so if multiple nations offer to fund pure longtermist goods in line with their citizens' AWTP.

Here is a final point to consider. One might think that it is true only *on the current margin* and *in public* that longtermists should push governments to adopt a catastrophe policy guided by cost-benefit analysis and altruistic willingness to pay. Once all the interventions justified by CBA-plus-AWTP have been funded, longtermists should lobby for even more government spending on preventing catastrophes. And in the meantime, longtermists should in private advocate for governments to fund existential-risk-reducing interventions that go beyond CBA-plus-AWTP.

We disagree. Longtermists can try to increase government funding for catastrophe-prevention by making longtermist arguments and thereby increasing citizens' AWTP, but they should not urge governments to depart from a CBA-plus-AWTP catastrophe policy. On the contrary, longtermists should as far as possible commit themselves to acting in accordance with a CBA-plus-AWTP policy in the political sphere. One reason why is simple: longtermists have moral reasons to respect the preferences of their fellow citizens.

To see another reason why, note first that longtermists working to improve government catastrophe policy could be a win-win. The present generation benefits because longtermists solve the collective action problem: they work to implement interventions that cost-effectively reduce everyone's risk of dying in a catastrophe. Future generations benefit because these interventions also reduce existential risk. But as it stands the present generation may worry that longtermists would go too far. If granted imperfectly accountable power, longtermists might try to use the machinery of government to place burdens on the present generation for the sake of further benefits to future generations. These worries may lead to the marginalization of longtermism, and thus an outcome that is worse for both present and future generations.

The best solution is compromise and commitment.[41] A CBA-plus-AWTP policy—founded as it is on citizens' preferences—is acceptable to a broad coalition of people. As a result, longtermists committing to act in accordance with a CBA-plus-AWTP policy make possible an arrangement that is significantly better than the status quo, both by longtermist lights and by the lights of the present generation. It also gives rise to other benefits of cooperation. For example, it helps to avoid needless conflicts in which groups lobby for opposing

[41] In this respect, the situation is analogous to Parfit's (1984: 7) hitchhiker case.

policies, with some substantial portion of the resources that they spend canceling each other out (see Ord 2015: 120–121, 135). With a CBA-plus-AWTP policy in place, those resources can instead be spent on interventions that are appealing to all sides.

There are many ways in which longtermists can increase and demonstrate their commitment to this kind of win-win compromise policy. They can speak in favor of it now, and act in accordance with it in the political sphere. They can also support efforts to embed a CBA-plus-AWTP criterion into government decision-making—through executive orders, regulatory statutes, and law—thereby ensuring that governments spend neither too much nor too little on benefits to future generations. Longtermists can also earn a reputation for cooperating well with others, by supporting interventions and institutions that are appealing to a broad range of people. In doing so, longtermists make possible a form of cooperation which is substantially beneficial to both the present generation and the long-term future.

7 Conclusion

Governments should be spending much more on averting threats from nuclear war, engineered pandemics, and AI. This conclusion follows from standard cost-benefit analysis. We need not assume longtermism, or even that future generations matter. In fact, even entirely self-interested Americans have reason to hope that the U.S. government adopts a catastrophe policy guided by cost-benefit analysis.

Longtermists should push for a similar goal: a government catastrophe policy guided by cost-benefit analysis and citizens' altruistic willingness to pay. This policy is achievable and democratically acceptable. It would also reduce existential risk by almost as much as a strong longtermist policy. This is especially so if longtermists succeed in making the long-term future a major moral priority of our time and if citizens' altruistic willingness to pay for benefits to the long-term future increases commensurately. Longtermists should commit to acting in accordance with a CBA-plus-AWTP policy in the political sphere. This commitment would help bring about a catastrophe policy that is much better than the status quo, for the present generation and long-term future alike.[42]

References

Aldy, J. E. and Viscusi, W. K. (2008), 'Adjusting the Value of a Statistical Life for Age and Cohort Effects', in *Review of Economics and Statistics* 90/3: 573–581.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016), 'Concrete Problems in AI Safety', *arXiv*, http://arxiv.org/abs/1606.06565

Armstrong, S., Bostrom, N., and Shulman, C. (2016), 'Racing to the Precipice: A Model of Artificial Intelligence Development', in *AI & Society* 31/2: 201–206.

Avin, S. and Amadae, S. M. (2019), 'Autonomy and Machine Learning as Risk Factors at the Interface of Nuclear Weapons, Computers and People', in V. Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk* (vol. 1, Euro-Atlantic Perspectives): 105–118.

[42] For helpful comments, we thank Mackenzie Arnold, David Denkenberger, Tomi Francis, Jakob Graabak, Samuel Hilton, Hannah Lovell, Toby Ord, Andreas Schmidt, Philip Trammell, Risto Uuk, Nikhil Venkatesh, an anonymous reviewer for Oxford University Press, and audience members at the 10th Oxford Workshop on Global Priorities Research.

Baum, S. D. (2015), 'The Far Future Argument for Confronting Catastrophic Threats to Humanity: Practical Significance and Alternatives', in *Futures* 72: 86–96.
Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future*, PhD thesis, Rutgers University, http://dx.doi.org/doi:10.7282/T35M649T
Beckstead, N. (2015), 'How Much Could Refuges Help Us Recover from a Global Catastrophe?', in *Futures* 72: 36–44.
Bipartisan Commission on Biodefense. (2021), *The Apollo Program for Biodefense: Winning the Race Against Biological Threats* (Bipartisan Commission on Biodefense), https://biodefensecommission.org/wp-content/uploads/2021/01/Apollo_report_final_v8_033121_web.pdf (accessed 24 February 2023).
Bipartisan Commission on Biodefense. (2022), *The Athena Agenda: Advancing the Apollo Program for Biodefense* (Bipartisan Commission on Biodefense), https://biodefensecommission.org/wp-content/uploads/2022/04/Athena-Report_v7.pdf (accessed 24 February 2023).
Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4/1: 15–31.
Bostrom, N., Douglas, T., and Sandberg, A. (2016), 'The Unilateralist's Curse and the Case for a Principle of Conformity', in *Social Epistemology* 30/4: 350–371.
Carlsmith, J. (2021), 'Is Power-Seeking AI an Existential Risk?', *arXiv*, http://arxiv.org/abs/2206.13353
Carlsmith, J. (2023), 'Scheming AIs: Will AIs fake alignment during training in order to get power?', *arXiv*, https://doi.org/10.48550/arXiv.2311.08379 (accessed 19 February 2025).
Caviola, L., Althaus, D., Mogensen, A. L., and Goodwin, G. P. (2022), 'Population Ethical Intuitions', in *Cognition* 218: 104941.
Centre for Long-Term Resilience. (2021), *Future Proof: The Opportunity to Transform the UK's Resilience to Extreme Risks* (Centre for Long-Term Resilience), https://11f95c32-710c-438b-903d-da4e18de8aaa.filesusr.com/ugd/e40baa_c64c0d7b430149a393236bf4d26cdfdd.pdf (accessed 24 February 2023).
Claxton, K., Ochalek, J., Revill, P., Rollinger, A., and Walker, D. (2016), 'Informing Decisions in Global Health: Cost Per DALY Thresholds and Health Opportunity Costs' (University of York Centre for Health Economics), https://www.york.ac.uk/media/che/documents/policybriefing/Cost%20per%20DALY%20thresholds.pdf (accessed 24 February 2023).
Cotra, A. (2020), 'Forecasting Transformative AI with Biological Anchors, Part 4: Timelines Estimates and Responses to Objections' (unpublished manuscript), https://docs.google.com/document/d/1cCJjzZaJ7ATbq8N2fvhmsDOUWdm7t3uSSXv6bD0E_GM (accessed 24 February 2023).
Cotra, A. (2022), 'Two-Year Update on My Personal AI Timelines', *AI Alignment Forum*, https://www.alignmentforum.org/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines (accessed 24 February 2023).
Coupe, J., Bardeen, C. G., Robock, A., and Toon, O. B. (2019), 'Nuclear Winter Responses to Nuclear War Between the United States and Russia in the Whole Atmosphere Community Climate Model Version 4 and the Goddard Institute for Space Studies ModelE', in *Journal of Geophysical Research: Atmospheres* 124/15: 8522–8543.
Cutler, D. M. and Summers, L. H. (2020), 'The COVID-19 Pandemic and the $16 Trillion Virus', in *Journal of the American Medical Association* 324/15: 1495–1496.
Dafoe, A. (2018), 'AI Governance: A Research Agenda'. (Future of Humanity Institute, University of Oxford), https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf (accessed 24 February 2023).
DeepMind. (2020), 'Specification Gaming: The Flip Side of AI Ingenuity', DeepMind, https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity (accessed 24 February 2023).
Favaloro, P. and Berger, A. (2021), 'Technical Updates to Our Global Health and Wellbeing Cause Prioritization Framework—Open Philanthropy', *Open Philanthropy*, https://www.openphilanthropy.org/research/technical-updates-to-our-global-health-and-wellbeing-cause-prioritization-framework/ (accessed 24 February 2023).
Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. (2018), 'When Will AI Exceed Human Performance? Evidence from AI Experts', in *Journal of Artificial Intelligence Research* 62, 729–754.
Graham, J. D. (2008), 'Saving Lives through Administrative Law and Economics', in *University of Pennsylvania Law Review* 157/2: 395–540.
Greaves, H. and MacAskill, W. (2021), 'The Case for Strong Longtermism', GPI Working Paper, No. 5-2021 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism/ (accessed 24 February 2023).
Hellman, M. E. (2008), 'Risk Analysis of Nuclear Deterrence', in *The Bent of Tau Beta Pi* 99/2: 14–22.
Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. (2022), 'Unsolved Problems in ML Safety', *arXiv*, http://arxiv.org/abs/2109.13916

Hirth, R. A., Chernew, M. E., Miller, E., Fendrick, A. M., and Weissert, W. G. (2000), 'Willingness to Pay for a Quality-Adjusted Life Year: In Search of a Standard', in *Medical Decision Making: An International Journal of the Society for Medical Decision Making* 20/3: 332–342.

Jebari, K. (2015), 'Existential Risks: Exploring a Robust Risk Reduction Strategy', in *Science and Engineering Ethics* 21/3: 541–554.

Kniesner, T. J. and Viscusi, W. K. (2019), 'The Value of a Statistical Life', in *Oxford Research Encyclopedia of Economics and Finance* (Oxford University Press), https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-138 (accessed 24 February 2023).

Krakovna, V. (2018), 'Specification Gaming Examples in AI', *Victoria Krakovna*, https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/ (accessed 24 February 2023).

Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., and Krueger, D. (2022), 'Goal Misgeneralization in Deep Reinforcement Learning', *arXiv*, http://arxiv.org/abs/2105.14111

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., Cheney, N., Chrabaszcz, P., Cully, A., Doncieux, S., Dyer, F. C., Ellefsen, K. O., Feldt, R., Fischer, S., Forrest, S., . . . Yosinski, J. (2020), 'The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities', in *Artificial Life* 26/2: 274–306.

Library of Congress (2022a), 'S.3799—PR EVENT Pandemics Act', https://www.congress.gov/bill/117th-congress/senate-bill/3799 (accessed 24 February 2023).

Library of Congress (2022b), S.4488 - Global Catastrophic Risk Management Act of 2022, https://www.congress.gov/bill/117th-congress/senate-bill/4488 (accessed 24 February 2023).

Library of Congress (2022c), (H.R.5376—Inflation Reduction Act of 2022) https://www.congress.gov/bill/117th-congress/house-bill/5376 (accessed 24 February 2023).

MacAskill, W. (2022), *What We Owe The Future: A Million-Year View* (Oneworld).

Matheny, J. G. (2007), 'Reducing the Risk of Human Extinction', in *Risk Analysis* 27/5: 1335–1344.

Metaculus. (2025a), 'By 2100, Will the Human Population Decrease by at Least 10% during Any Period of 5 Years?', *Metaculus*, https://www.metaculus.com/questions/1493/global-population-decline-10-by-2100/ (accessed 7 January 2025).

Metaculus. (2025b), 'If a Global Catastrophe Occurs, Will it Be Due to Biotechnology or Bioengineered Organisms?', *Metaculus*, https://www.metaculus.com/questions/1502/ragnar%25C3%25B6k-question-series-if-a-global-catastrophe-occurs-will-it-be-due-to-biotechnology-or-bioengineered-organisms/ (accessed 7 January 2025).

Metaculus. (2025c), 'Will There Be a Global Thermonuclear War by 2070?', *Metaculus*, https://www.metaculus.com/questions/3517/will-there-be-a-global-thermonuclear-war-by-2070/ (accessed 7 January 2025).

Michigan, et al. v. Environmental Protection Agency, et al. (No. 14-46); Utility Air Regulatory Group v. Environmental Protection Agency, et al. (No. 14-47); National Mining Association v. Environmental Protection Agency, et al. (No. 14-49), No. 14-46, (2015), (135 Supreme Court of the United States 2699 29 July 2015).

Millett, P. and Snyder-Beattie, A. (2017), 'Existential Risk and Cost-Effective Biosecurity', in *Health Security* 15/4: 373–383.

Mills, M. J., Toon, O. B., Lee-Taylor, J. M., and Robock, A. (2014), 'Multi-Decadal Global Cooling and Unprecedented Ozone Loss Following a Regional Nuclear Conflict', in *Earth's Future 2*/4: 161–176.

Muehlhauser, L. (2021), 'Treacherous Turns in the Wild', *Luke Muehlhauser,* https://lukemuehlhauser.com/treacherous-turns-in-the-wild/ (accessed 24 February 2023).

Murphy, T. (2013), 'The First Level of Super Mario Bros is Easy with Lexicographic Orderings and Time Travel', http://www.cs.cmu.edu/~tom7/mario/mario.pdf (accessed 24 February 2023).

National Science Foundation. (2023), 'Safe Learning-Enabled Systems', *National Science Foundation*, https://beta.nsf.gov/funding/opportunities/safe-learning-enabled-systems (accessed 24 February 2023).

Neumann, P. J., Cohen, J. T., and Weinstein, M. C. (2014), 'Updating Cost-Effectiveness—The Curious Resilience of the $50,000-per-QALY Threshold', in *The New England Journal of Medicine* 371/9: 796–797.

Ngo, R., Chan, L. and Mindermann, S. (2024), 'The Alignment Problem from a Deep Learning Perspective', in *The Twelfth International Conference on Learning Representations*, https://openreview.net/forum?id=fh8EYKFKns (accessed 19 February 2025).

Nuclear Threat Initiative. (2020a), *Preventing the Next Global Biological Catastrophe* (Agenda for the Next Administration: Biosecurity), https://media.nti.org/documents/Preventing_the_Next_Global_Biological_Catastrophe.pdf (accessed 24 February 2023).

Nuclear Threat Initiative. (2020b), *Reducing Nuclear Risks: An Urgent Agenda for 2021 and Beyond* (Agenda for the Next Administration: Nuclear Policy), https://media.nti.org/documents/Reducing_Nuclear_Risks_An_Urgent_Agenda_for_2021_and_Beyond.pdf (accessed 24 February 2023).

OpenAI. (2017), 'Learning from Human Preferences', *OpenAI,* https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/ (accessed 24 February 2023).

Ord, T. (2015), 'Moral Trade', in *Ethics* 126/1: 118–138.

Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).

Our World in Data. (2019), 'Number of Deaths by Cause, United States, 2019', *Our World in Data*, https://ourworldindata.org/grapher/annual-number-of-deaths-by-cause?country=~USA (accessed 24 February 2023).

Our World in Data. (2025), 'Daily and Total Confirmed COVID-19 Deaths, United States', *Our World in Data*, https://ourworldindata.org/grapher/total-daily-covid-deaths (accessed 7 January 2025).

Parfit, D. (1984), *Reasons and Persons* (Clarendon Press).

PopulationPyramid. (2024), 'Population Pyramid for the United States of America, 2035', *PopulationPyramid.net*, https://www.populationpyramid.net/united-states-of-america/2035/ (accessed 7 January 2025).

Posner, E. A. and Sunstein, C. R. (2017), 'Moral Commitments in Cost-Benefit Analysis', in *Virginia Law Review* 103: 1809–1860.

Posner, R. (2004), *Catastrophe: Risk and Response* (Oxford University Press).

Reisner, J., D'Angelo, G., Koo, E., Even, W., Hecht, M., Hunke, E., Comeau, D., Bos, R., and Cooley, J. (2018), 'Climate Impact of a Regional Nuclear Weapons Exchange: An Improved Assessment Based On Detailed Source Calculations', in *Journal of Geophysical Research: Atmospheres* 123/5: 2752–2772.

Robock, A., Oman, L., and Stenchikov, G. L. (2007), 'Nuclear Winter Revisited with a Modern Climate Model and Current Nuclear Arsenals: Still Catastrophic Consequences', in *Journal of Geophysical Research* 112/D13.

Rodriguez, L. (2019a), 'How Bad Would Nuclear Winter Caused by a US-Russia Nuclear Exchange Be?', *Rethink Priorities*, https://rethinkpriorities.org/publications/how-bad-would-nuclear-winter-caused-by-a-us-russia-nuclear-exchange-be (accessed 24 February 2023).

Rodriguez, L. (2019b), 'How Likely Is a Nuclear Exchange between the US and Russia?', *Rethink Priorities*, https://rethinkpriorities.org/publications/how-likely-is-a-nuclear-exchange-between-the-us-and-russia (accessed 24 February 2023).

Sandberg, A. and Bostrom, N. (2008), 'Global Catastrophic Risks Survey', Technical Report #2008-1 (Future of Humanity Institute, Oxford University), https://www.fhi.ox.ac.uk/reports/2008-1.pdf (accessed 24 February 2023).

Schubert, S., Caviola, L., and Faber, N. S. (2019), 'The Psychology of Existential Risk: Moral Judgments about Human Extinction', in *Scientific Reports* 9/1: 15100.

Seitz, R. (2011), 'Nuclear Winter Was and Is Debatable', in *Nature* 475/7354: 37.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. (2022), 'Compute Trends Across Three Eras of Machine Learning', *arXiv*, http://arxiv.org/abs/2202.05924

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. (2022), 'Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals', *arXiv*, http://arxiv.org/abs/2210.01790

Soares, N., Fallenstein, B., Yudkowsky, E., and Armstrong, S. (2015), 'Corrigibility', in *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*, https://cdn.aaai.org/ocs/ws/ws0067/10124-45900-1-PB.pdf (accessed 19 February 2025).

Steinhardt, J. (2022), 'AI Forecasting: One Year In', *Bounded Regret*, https://bounded-regret.ghost.io/ai-forecasting-one-year-in/ (accessed 24 February 2023).

Stein-Perlman, Z., Weinstein-Raun, B., and Grace, K. (2022), '2022 Expert Survey on Progress in AI', *AI Impacts*, https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/ (accessed 24 February 2023).

Sunstein, C. R. (2020), How Change Happens (MIT Press).

Teran, N. (2022), 'Preventing Pandemics Requires Funding', *Institute for Progress*, https://progress.institute/preventing-pandemics-requires-funding/ (accessed 24 February 2023).

The Nucleic Acid Observatory Consortium. (2021), 'A Global Nucleic Acid Observatory for Biodefense and Planetary Health'. *arXiv*. http://arxiv.org/abs/2108.02678

The U.S. National Archives and Records Administration (2012), Executive Order No. 13,563, Code of Federal Regulations, Title 3 215, https://www.archives.gov/federal-register/codification/executive-order/12291.html (accessed 24 February 2023).

The White House. (2022a), 'A Return to Science: Evidence-Based Estimates of the Benefits of Reducing Climate Pollution', *The White House*, https://www.whitehouse.gov/cea/written-materials/2021/02/26/

a-return-to-science-evidence-based-estimates-of-the-benefits-of-reducing-climate-pollution/ (accessed 24 February 2023).
The White House. (2022b), 'Joint Statement of the Leaders of the Five Nuclear-Weapon States on Preventing Nuclear War and Avoiding Arms Races', *The White House*, https://www.whitehouse.gov/briefing-room/statements-releases/2022/01/03/p5-statement-on-preventing-nuclear-war-and-avoiding-arms-races/ (accessed 24 February 2023).
The White House. (2022c), 'The Biden Administration's Historic Investment in Pandemic Preparedness and Biodefense in the FY 2023 President's Budget', *The White House*, https://www.whitehouse.gov/briefing-room/statements-releases/2022/03/28/fact-sheet-the-biden-administrations-historic-investment-in-pandemic-preparedness-and-biodefense-in-the-fy-2023-presidents-budget/ (accessed 24 February 2023).
Thornley, E. (2022), 'Calculating Expected American Life-Years Saved by Averting a Catastrophe in 2033' (unpublished manuscript), https://docs.google.com/spreadsheets/d/1mguFYc06mw2Bdv85Viw6CQq4mt-mqPdiU4mQRi1s0Yo (accessed 24 February 2023).
U.K. Treasury. (2003), *The Green Book: Appraisal and Evaluation in Central Government*, (TSO), https://webarchive.nationalarchives.gov.uk/ukgwa/20080305121602/http://www.hm-treasury.gov.uk/media/3/F/green_book_260907.pdf (accessed 24 February 2023).
U.S. Department of Transportation. (2021a), *Departmental Guidance on Valuation of a Statistical Life in Economic Analysis* (U.S. Department of Transportation), https://www.transportation.gov/office-policy/transportation-policy/revised-departmental-guidance-on-valuation-of-a-statistical-life-in-economic-analysis (accessed 24 February 2023).
U.S. Department of Transportation. (2021b), *Departmental Guidance: Treatment of the Value of Preventing Fatalities and Injuries in Preparing Economic Analyses*, https://www.transportation.gov/sites/dot.gov/files/2021-03/DOT%20VSL%20Guidance%20-%202021%20Update.pdf (accessed 24 February 2023).
U.S. Environmental Protection Agency. (2010), *Valuing Mortality Risk Reductions for Environmental Policy: A White Paper (2010)* (U.S. Environmental Protection Agency), https://www.epa.gov/sites/default/files/2017-08/documents/ee-0563-1.pdf (accessed 24 February 2023).
U.S. Government (1982), Executive Order No. 12,291, Code of Federal Regulations, Title 3 127, https://www.govinfo.gov/app/details/CFR-2012-title3-vol1/CFR-2012-title3-vol1-eo13563/summary (accessed 24 February 2023).
U.S. Government (1989), 'Ohio v. U.S. Dept. Of the Interior, 880 F. 2d 432' (Court of Appeals, Dist. of Columbia Circuit 1989).
U.S. Government (2010), 'Nondiscrimination on the Basis of Disability in State and Local Government Services, 75 Federal Register 56164 (Sept. 15, 2010) (codified at 28 Code of Federal Regulations, pt. 35)'.
U.S. Government (2012a), 'National Standards to Prevent, Detect, and Respond to Prison Rape, 77 Federal Register 37106 (June 20, 2012) (codified at 28 Code of Federal Regulations, pt. 115)'.
U.S. Government (2012b), 'Scope of Review, 5 U.S. Code §706(2)(A)', https://www.govinfo.gov/app/details/USCODE-2011-title5/USCODE-2011-title5-partI-chap7-sec706/summary (accessed 24 February 2023).
U.S. National Security Commission on Artificial Intelligence. (2021), *Final Report* (U.S. National Security Commission on Artificial Intelligence), https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf (accessed 24 February 2023).
U.S. Office of Management and Budget. (2022), *Budget of the U.S. Government: Fiscal Year 2023* (U.S. Office of Management and Budget) https://www.whitehouse.gov/wp-content/uploads/2022/03/budget_fy2023.pdf (accessed 8 April 2022).
U.S. Social Security Administration. (2022), *Actuarial Life Table* (Social Security Administration), https://www.ssa.gov/oact/STATS/table4c6.html (accessed 24 February 2023).
Vinding, M. (2020), *Suffering-Focused Ethics: Defense and Implications* (Ratio Ethica).
Watson, C., Watson, M., Gastfriend, D., and Sell, T. K. (2018), 'Federal Funding for Health Security in FY2019', in *Health Security* 16/5: 281–303.
Wiblin, R. and Ord, T. (2020), 'Toby Ord on *The Precipice* and Humanity's Potential Futures', The 80,000 Hours Podcast with Rob Wiblin *(80,000 Hours)*, https://80000hours.org/podcast/episodes/toby-ord-the-precipice-existential-risk-future-humanity/ (accessed 24 February 2023).
Wiener, J. B. (2016), 'The Tragedy of the Uncommons: On the Politics of Apocalypse', in *Global Policy* 7/S1: 67–80.
Xia, L., Robock, A., Scherrer, K., Harrison, C. S., Bodirsky, B. L., Weindl, I., Jägermeyr, J., Bardeen, C. G., Toon, O. B., and Heneghan, R. (2022), 'Global Food Insecurity and Famine from Reduced Crop, Marine Fishery and Livestock Production due to Climate Disruption from Nuclear War Soot Injection', in *Nature Food* 3/8: 586–596.

17
Longtermist Myopia

Amanda Askell and Sven Neth

1 Introduction

Longtermists have argued that positively influencing the long-term future should be a key moral priority today (MacAskill 2022: 11). But how should we weigh improvements to the long-term future against improvements to the present? When it comes to this question, longtermists face possible objections from two sides. If longtermists say that we should give the same weight to present improvements relative to the long-term future as 'neartermists' (non-longtermists) do, it's not clear what longtermism really adds to our moral decision-making. Let's call this the objection from irrelevance. On the other hand, if longtermists say that we should be willing to neglect present people in favor of improving the lives of far-future people, their theory conflicts with common moral intuitions. Let's call this the objection from future fanaticism.[1]

In this chapter, we argue that longtermism represents a less radical departure from neartermism than one might expect. In section 2, we introduce the concepts of 'myopia' and 'hyperopia'. These are the degrees to which an agent's moral choices approximate or deviate from those of a neartermist agent. In section 3, we outline several considerations that cause longtermist agents to be hyperopic and make decisions in ways that deviate strongly from their neartermist counterparts. In section 4, we show that several important considerations push longtermists towards myopia instead, and will lead them to make moral choices that approximate those of neartermist agents. We argue that even if an agent believes that the future matters just as much as the present from a moral point of view, there are strong reasons for the agent to focus more on the near-term consequences of their actions.

We do not claim that the considerations in favor of myopia dissolve the differences between longtermism and neartermism. Longtermists will often still act differently from neartermists even if reasons for myopia are taken into account. However, we believe that the gap between longtermist and neartermist decision makers is smaller than one might initially expect, and the practical consequences of accepting longtermism are less revisionary than they might initially appear.

It is not clear whether this chapter is, on the whole, a friend or foe of longtermism. We argue that adopting longtermism will not change the behavior of neartermist agents as much as one might think, which constitutes at least a weak form of the objection from irrelevance. But this may also help longtermists avoid the most damning versions of the

[1] We call this 'future fanaticism' because of its close relation to concerns about 'fanaticism' that arise when agents use expected value reasoning in cases that involve very small probabilities of creating very large amounts of value (Monton 2019; Beckstead and Thomas 2021; Wilkinson 2022).

Amanda Askell and Sven Neth, *Longtermist Myopia* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Amanda Askell and Sven Neth 2025.
DOI: 10.1093/9780191979972.003.0017

objection from future fanaticism without retreating to complete irrelevance. So we are not entirely sure ourselves whether this constitutes an attack on or defense of longtermism, and leave it for the reader to decide.

2 Longtermism and myopia

According to *strong longtermism*, the value of an action is driven primarily by its expected impact on the far future. The axiological formulation of strong longtermism says that, when it comes to the most important decisions we face:

(i) Every option that is near-best overall is near-best for the far future.
(ii) Every option that is near-best overall delivers much larger benefits in the far future than in the near future [relative to the status quo]. (Greaves and MacAskill 2022: 3)

Similarly, the deontic version of strong longtermism says that, when it comes to the most important decisions we face:

(i) One ought to choose an option that is near-best for the far future.
(ii) One ought to choose an option that delivers much larger benefits in the far future than in the near future [relative to the status quo]. (Greaves and MacAskill 2022: 26)

Both formulations of strong longtermism presuppose that, in the most important decisions we face today, at least one option available to us delivers much larger absolute benefits in the far future than it does in the near future.[2]

A common view used to support views like strong longtermism is what we might call 'time indifference': the view that the consequences of our actions do not matter less just because they occur in the future. On this view, we should not intrinsically discount the moral value of people and events merely because they are far away in time, just as many of us believe that we should not intrinsically discount the moral value of people and events merely because they are far away in space (Beckstead 2013; Ord 2020; Greaves and MacAskill 2021; MacAskill 2022).

Although time indifference is a view often cited in support of longtermism, longtermism is a normative conclusion that can fall out of many different normative views, in the same way that 'killing is almost always wrong' is a normative conclusion that can fall out of many different normative views. Longtermism can be supported by any normative view that gives a great deal of moral importance to the far future, of which time indifference is just one example.

Throughout this chapter, we will use 'longtermist agents' to refer to agents who accept strong longtermism as it is formulated above. We will model such agents as having a pure temporal discount rate of zero. We will use 'neartermist agents' to refer to agents who believe

[2] Suppose we discovered with certainty that the world was going to end tomorrow. Condition (ii) of strong axiological and deontic longtermism would no longer be true of any action in our option set, since none of the options available to us would deliver larger absolute benefits in the far future than they do in the near future. Strong longtermism is not intended to apply to such scenarios, however.

their actions should be guided primarily by the impact they will have in the immediate future. This means that, even if a neartermist agent was omniscient about the long-term impact of their actions, they would still give priority to their impact on the immediate future. We will model such agents as having some strictly positive pure temporal discount rate.

The impact of actions on the long-term future will almost always be a critical factor in the moral decisions of longtermist agents, but it won't always be the decisive factor. To see why, suppose you are a longtermist and the only actions you can take are to send financial assistance to country A, send financial assistance to country B, or waste the money. You know country A would benefit more from the assistance than country B in the near term. You also know that the long-term impact of giving assistance to country A or country B will be large and positive relative to wasting the money, but the impact will be similar regardless of which country you send assistance to. So you decide to send assistance to country A, since it will benefit more from it in the near term. In this choice situation, it seems that the near-term impact of the assistance is the 'decisive factor' in your choice to send the money to country A rather than country B.[3]

If we accept strong longtermism, we must first look at the long-term impact of the actions available to us and eliminate any actions whose long-term impact is worse than that of at least one other action to a degree that could not plausibly be outweighed by near-term value or by other ethical considerations. Once we have eliminated all the actions whose long-term impact is worse than at least one other action to the relevant degree, we can choose between any remaining actions using other factors such as immediate impact without violating any of the principles of strong longtermism as formulated above.

It is now helpful to introduce the concepts of 'myopia' and 'hyperopia'. We will say that an agent is behaving *myopically* if (and to the degree that) her moral choices approximate those that an agent with a strictly positive temporal discount rate would make in the same situation. We will say an agent is behaving *hyperopically* if (and to the degree that) her moral choices deviate from those that an agent with a strictly positive temporal discount rate would make in the same situation. The higher the discount rate that the agent's choices approximate, the more myopically she is behaving. Myopia and hyperopia are both relative to decision scenarios and not agents: the same agent could behave myopically in one decision scenario and hyperopically in a different decision scenario.

Longtermist agents will generally act less myopically than their neartermist counterparts since they give more moral weight to the long-term consequences of their actions. But longtermist agents will display varying degrees of myopia across decision situations. Consider the decision above, in which you had to choose where to send financial assistance. In this scenario, a neartermist and a longtermist would both probably rank sending the money to country A above sending the money to country B above wasting the money.

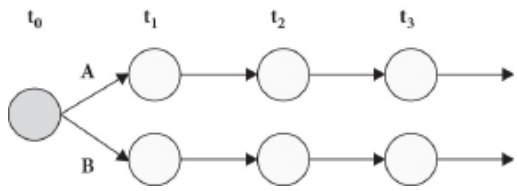
[3] We can use this case to construct a possible objection to strong longtermism. Suppose wasting the money was removed as a possible option and therefore could not be treated as the 'status quo' (assuming the status quo must be in the agent's option set). We would instead have to treat 'sending assistance to country A' or 'sending assistance to country B' as the status quo. But this means there is no near-best option available to the agent according to axiological strong longtermism, since none of the options would deliver much larger benefits in the far future than in the near future relative to the status quo. So when wasting money is an option (C), we can rank the actions $A \succ B \succ C$. But if C is removed then $A \not\succ B$ and $B \not\succ A$. This seems in conflict with the independence of irrelevant alternatives (IIA). We believe such an objection should not concern the strong longtermist, however. Strong longtermism is a hypothesis about the important decisions agents face today and is not intended to apply to all possible decision situations. The strong longtermist can therefore simply respond that the scenario in which the option of wasting the money is removed does not reflect the kind of decisions facing agents today.

So in this particular decision situation, the longtermist will behave as myopically as the neartermist.

In the next section, we explore the ways in which longtermism tends towards hyperopia. We outline features that longtermists will prioritize relative to neartermists, and outline some of the major differences between neartermist and longtermist moral decision-making.

3 Longtermist hyperopia

Suppose you've just finished graduate school and are deciding whether to apply for academic jobs (A) or pursue a promising position outside academia (B). You want to pick the option that will have the best impact on your future life. This future life can be broken down into a series of time steps, and you want to know how good each of these series would be conditional on taking the academic or non-academic role. If an oracle were helping you with this decision, they could let you look at the entire path of your life unfolding from this point in time onward.[4]

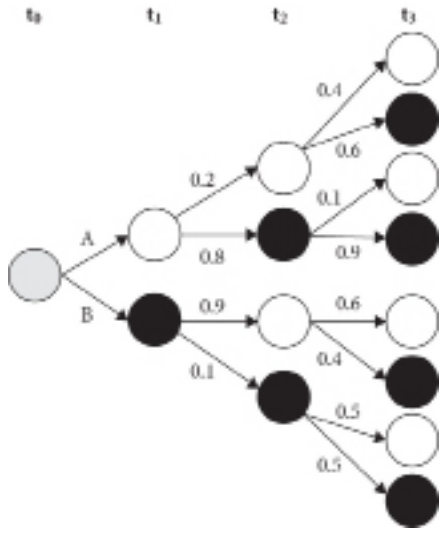


If you could see these futures, one might be clearly better than the other. For example, in one of them you eventually become wealthy and live for a long time, while in the other you meet an untimely death. The oracle would be very helpful if this were the case, since the key events determining which path you prefer could be extremely hard to predict from your current vantage point. Perhaps your untimely death occurs because you have an accident on a holiday four years after you decide to apply for academic jobs, while the same holiday does not occur if you had left academia. Without the oracle, the best you can do is consider all the different possible futures following your decision and the subjective likelihood of each of these possibilities.[5]

In the real world, the subjectively possible consequences of our actions will be vast. For now, we will simplify things by treating the future as a series of discrete time steps and supposing that each time step always leads to at most two possible states: one good and one bad. You assign a value of 1 to good states and 0 to bad states (i.e. they are all good and bad to the same degree). We can then represent the possible outcomes of your choice between A and B at three future time steps on a graph, where white circles represent bad states and black circles represent good ones.

[4] To simplify things, we assume time is deterministic and linear after you make your decision. Indeed, we employ a fairly naive account of time throughout this chapter. We acknowledge that an appropriate account of temporal discounting should be compatible with—and ideally explained in terms of—theories of space and time considered plausible within physics.

[5] We assume that subjective uncertainty should be modeled with a precise probability function (Titelbaum 2019) and set aside imprecise probabilities (Mahtani 2019).



Let's use 'direct value' to refer to the direct value of the state, and 'indirect value' to refer to the value of all the future states this state can lead to weighted by their subjective likelihood. In the graph above, the direct value of A is 0 and the direct value of B is 1. But if we multiply the value of each subsequent state by its subjective probability, we can see that the indirect value of A is 1.64 while the indirect value of B is 0.51.[6] So if there is no time step after t_3, the total value of A is 1.64 while the total value of B is 1.51.[7] Choosing A leads us to a worse state in the immediate term but is the better choice once we consider the entire future.[8]

Here we have modeled the decision as a simple Partially Observable Markov Decision Processes (POMDP).[9] In a POMDP, there are discrete states and conditional probabilistic transitions between those states. These are the time steps and the rational subjective likelihood that one step will lead to another. The agent has actions she can take in some of the states, which causes the state to transition to another with some probability. At the point of decision between A and B, we assume for simplicity that it will cause a transition into one of two future states with probability 1. There is also a reward at each state, which we have assumed to be either 1 or 0 for every future time step. Note that the agent might also

[6] Since we assume no discounting, the indirect value of A is 0.8 + (0.8 * 0.9) + (0.2 * 0.6) and the indirect value of B is 0.1 + (0.1 * 0.5) + (0.9 * 0.4).

[7] We compute the total value of A and B using the Bellman Equation, which equates the total value of a state with the direct value of the state plus the discounted value of each successor state weighted by its probability. Here is the Bellman Equation for the value of state s when following policy π, where π(s) is the action recommended by policy π in state s, R(s, π(s)) is the reward for action π(s) in state s, S is the set of all states, and γ is the discount factor: $V(\pi, s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi(s)) V(\pi, s')$ (Puterman 1994). This assumes that we know which state we are in. If the state is uncertain, we can just consider our belief states and apply the Bellman Equation to these states (Russell and Norvig 2018: Ch. 17). In our example, the discount rate is zero.

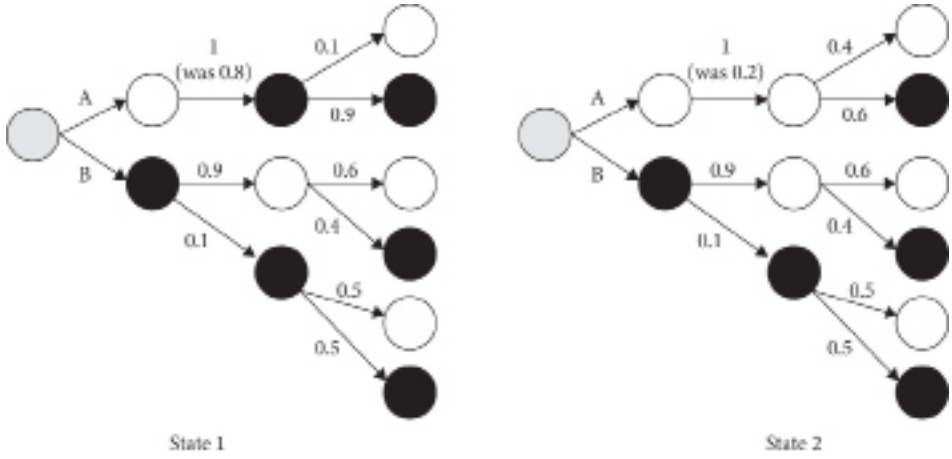
[8] We assume a subjective consequentialist background theory on which you should choose an action which maximizes expected value (Savage 1954).

[9] Kaelbling, Littman, and Cassandra (1998) and Russell and Norvig (2018: ch. 17.4) provide overviews of POMDPs. In most examples below, we assume a fixed finite time horizon. In general, there are problems with infinite-horizon POMDPS, especially if we assume that there is no discounting. For example, it has been shown that if we assume an infinite time horizon, the problem of finding an optimal policy is undecidable (Madani, Hanks, and Condon 2003). We do not address the problems with infinite POMDPs in this chapter, but we do believe that such problems are relevant to longtermist decision makers.

be uncertain about the true state, which will be important for the example in the next section. Finally, there is a discount rate $\gamma \in [0, 1]$ across future rewards, which we will return to below. For now we assume that the discount rate is zero, since we are interested in modeling a decision maker that does not intrinsically discount the future.

3.1 The value of information

One of the advantages of POMDPs is that they let us quantify the value of getting information about present or future states. Suppose that in the example above we could pay some units of value to achieve certainty about what the state immediately following a choice of A would be:



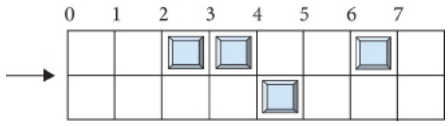
If the future state were revealed to be State 1, the value of A would be 1.9. If the future state were revealed to be State 2, the value of A would be 0.6. In both states the value of B would continue to be 1.51. Without learning the information, you should choose option A, since the expected value of A based on your prior information is 1.64, which is greater than the expected value of B. If you learn you are in State 1, you will choose A since 1.9 > 1.51. But if you learn you are in State 2, you will choose B since 1.51 > 0.6. Since you assign prior probability 0.8 to State 1 and 0.2 to State 2, the expected value of making your decision after learning which state you are in is $0.8 * 1.9 + 0.2 * 1.51 = 1.822$. So you should be willing to pay up to $1.822 - 1.64 = 0.182$ units of value to learn which state you are in. This is the value of that information.[10]

It's worth emphasizing that information can be valuable even if it doesn't tell you exactly which state you are in, but instead gives you a noisy signal which shifts your subjective probabilities. Imagine you have a credence of 0.5 there is $100 in a box to your right and a credence of 0.5 it's in the box to your left. You can get an imperfect hint about which box the money is in that will increase your credence in one of the boxes from 0.5 to 0.6 and

[10] The value of information is discussed by Raiffa and Schlaifer (1961), Good (1967), and many others. Kadane, Schervish, and Seidenfeld (2008) note that there are some circumstances in which decision makers will reject free information, including imprecise credences and merely finitely additive subjective probabilities. Furthermore, alternatives to expected utility theory will sometimes advise agents to reject free information (Buchak 2010). We will set these complications aside.

decrease the other box from 0.5 to 0.4. How much should you be willing to pay for this information? Since the difference in expected value between your options will go from $0 without the hint to $20 with the hint, you should be willing to pay up to $20 for the hint.

Actions that reveal information about future states are sometimes more valuable than actions that maximize expected direct value relative to current information. Suppose an agent is playing a game that involves advancing across an eight-tile-long corridor. At each stage they can choose to advance via the top tile or the bottom tile without searching, or they can pay a penalty of 0.5 units of value to search for traps in front of them with 100% success and advance immediately after searching. There is a 50% chance that a trap will spawn on at most one of the next steps every time they advance, meaning there will always be at least one untrapped square before them, but if they choose randomly there's a 25% chance they'll step onto a trap. If they do move onto a square with a trap on it, the game is over. If they move onto an untrapped tile they get 1 unit of value. So, at the end of the game, once all the traps have been revealed, the board might look as follows:



Suppose an agent wants to pay to search only if doing so will maximize expected value for her and she has a discount rate of zero. At each stage, her choice is between searching and advancing and having a guarantee of 0.5 units of value, or advancing without search and having a 75% chance of 1 unit of value and a 25% chance of no future value at all. So if an agent is deciding what to do at time step n and we let r be the expected value of all time steps after n, the value of searching and advancing ('search') is $r + 0.5$, while the value of advancing without search ('choose') is $0.75r + 0.75$.[11]

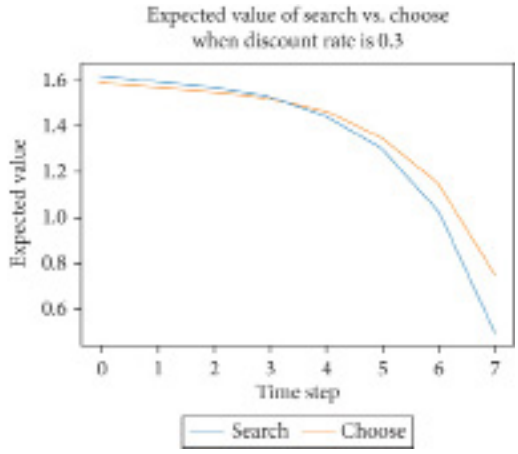
The agent can work out what she should do in this case using backward induction: starting from the final step and calculating the expected total value of searching or not searching on the next turn.[12] Below is the expected value of searching and not searching at each turn.

[11] In this case, 0.5 and 0.75 are the expected direct value of the next step conditional on searching and not searching in this round respectively, while r and $0.75r$ are the expected value of each step after the immediate next step conditional on searching and not searching in this round respectively.

[12] Bertsekas (2005) provides a detailed overview of how to solve such optimization problems.

This means that the agent should pay to search until she is on step 6, at which point she should pick the left or right tile at random. If we wanted to compare different policies, the expected value of an action at each step is calculated relative to the policy the agent will use at future steps.[13] In this case, it is easy to show that the expected value of searching until step 6 is better than the expected value of policies which recommend a higher or lower rate of searching.

As we have seen, the value of information is the difference between the value of the actions we will take conditional on the information compared with those we will take without it. The more some information can improve our prospects, the more valuable it is. Given this, longtermists will tend to assign very high value to information which is relevant to the impact of their actions on the long-term future. Such information will be less valuable to an agent who discounts the future. If an agent with a discount rate of 0.3 were playing the game above, for example, the agent would search until she is on step 4, at which point she will start to pick left or right at random.[14]



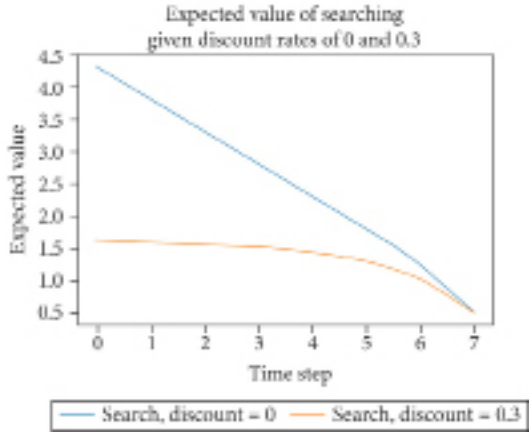
Discounting the future reduces the value of searching for traps because the expected value lost by ending the game is lower, since most of the value it generates happens in the future. This point holds more generally: information about the future is less valuable to you if you are discounting the future. So when we have the option to learn information about the future, whether one discounts the future can have a potentially large impact on how much one is willing to pay in order to learn the information. This is one way in which the preferences and actions of longtermists will deviate from those of neartermists.

[13] The expected value of the suboptimal action (searching or choosing) is calculated by assuming the agent deviates from her strategy for this round only and will do whatever is optimal at the next step.

[14] In other words, the agent values the very next state in accordance with its direct value, the state after that at 0.7 times its value, the state after that at 0.7^2 its value, and so on.

3.2 Option value

In the trap game above, an agent with a discount rate of zero places more value on getting information that prevents her from stepping on a trap. This is because she values the future states that would be lost if she steps on a trap more than someone who discounts those future states. The lower her discount rate and the longer the possible future, the more she will value actions that preserve the ability to access positive future states—in other words, the more value she places on *preserving options*. We can see this by comparing the expected value that agents with different discount rates attribute to searching for traps at each time step.



The expected value of searching is much higher for the agent with a discount rate of zero.[15] Although she is required to pay 0.5 units of value to search, she would be willing to pay much more than this, especially at the beginning of the game when the potential future value is higher. The game as a whole has more potential value for her, and therefore actions that keep her in the game are also more valuable.

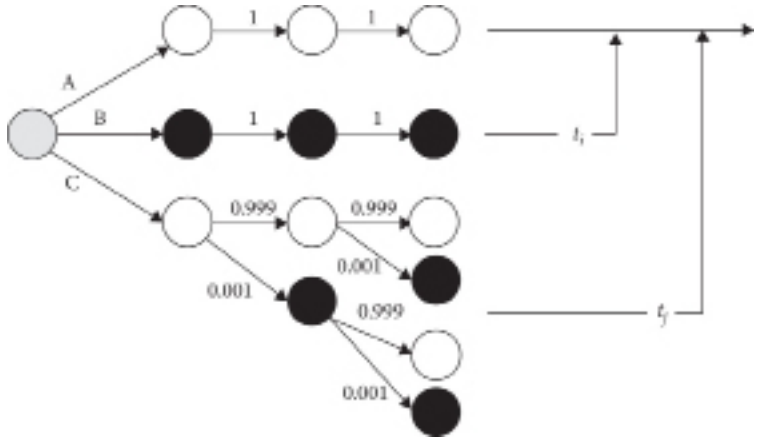
This suggests that when we have the option to take actions which preserve our future, agents who do not discount the future will value these actions much higher than agents who discount the future.[16] Most notably, this means that agents who do not discount the future will assign much higher value to preventing existential risks. These are risks that threaten to cause human extinction or the permanent reduction of humanity's future potential.[17] Preventing existential risks has indeed been one of the key issues of the longtermist agenda (Ord 2020).

[15] The expected value of choosing is also higher because the game as a whole has more value when the agent has a lower discount rate, and this is reflected in the value of both actions available to the agent.

[16] This assumes that preserving options is good in expectation: that the options that future people will explore are likely to create value rather than disvalue. If we expect future options to be used more for bad than good, preserving these future options could be very bad indeed.

[17] What about actions that cause a persistent gain in value across the future but do not do so via preserving future options, such as improving institutions that will persist into the far future? For longtermists, actions that increase the likelihood that future options can and will be used for good may be extremely valuable. See MacAskill (2022: ch. 4) on the topic of value lock-in, for example.

Even if the likelihood of positive future states is relatively low, if the future is sufficiently long, the difference between outcomes with no future value and outcomes that have a small but non-zero chance of positive value across future times can be vast. Consider an example in which there are three possible actions: A, B, and C. Action A leads to a state where there is no value at any future time. Action B leads to a state where there is a guarantee of positive future value for some very long but finite amount of time until t_i after which there is no more value (it intersects with the outcome of A at t_i). Action C leads to a state where there is a 99.9% chance of no value at the next step and a 0.01% chance of positive value at the next step until t_j.



Given a future discount rate of zero, action C is subjectively better than action B for any $j > 1000i$. Regardless of how long the period until t_i is, if the future is long enough then a small chance of good outcomes in the long term is better than a guarantee of good outcomes in the near term, even if the 'near-term' is very long indeed.

If we remove the constraint that the value of each outcome is either 0 or 1, we can construct analogous cases involving more immediate differences in value. Parfit (1984: 453) considers a thought experiment in which either all people in the world survive, only 1% of people survive, or all people in the world die. Although people might initially think that the 1% of people surviving and no one surviving are close in terms of their badness, if one does not discount the future then the badness of 1% of people surviving is much closer to the scenario in which no one dies than it is to the scenario in which everyone dies. If 1% of people survive, the future could still be full of valuable lives just as it is in the scenario in which no one dies. Such a future is impossible after an event that kills everyone.

Of course, actions that preserve option value are not limited to actions that directly avert catastrophes. They will include actions that improve the likelihood that future people will also care about averting existential catastrophes. Successfully averting existential risks via indirect means, such as improving institutions or increasing how much people care about avoiding them, will also be highly valuable from a longtermist perspective.[18] But it

[18] Beckstead (2013: 10) discusses the related distinction between 'targeted' and 'broad' interventions to shape the far future.

is clear that the high value placed on preserving options is another key difference between longtermists and neartermists. In general, longtermists will assign the most value to actions that allow for a persistent gain in value in the long-term, most plausibly by preventing existential catastrophe.[19] And they will assign much more value to these actions than their neartermist counterparts.

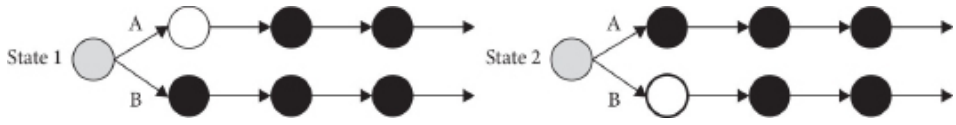
4 Longtermist myopia

We have shown that longtermists will often have different priorities from agents who discount the future. Longtermists will place much more value on information that can be used to influence the long-term future. They will also place much more value on preserving the long-term future and on options that reduce existential risks. We will now discuss considerations that may push longtermists to behave more myopically—considerations that may cause their actions to be closer to those of neartermists.

4.1 Causal diffusion as a reason for myopia

Causal diffusion is the idea that the causal consequences of our actions decay over time, just like ripples in a pond. The intuition here is that *lasting changes are rare*. Different jobs end up with similar long-term consequences in terms of value, ending one war rarely prevents future wars from taking place, and so on.[20] While our actions might make a causal difference in the short run, the world eventually goes back to a kind of 'default state' no matter what we do, and so the causal effects of our actions wash out.

For example, you might think that even if applying for academic jobs or pursuing a promising position outside academia will make a causal difference in the immediate future, it's unlikely to make a causal difference to the value of the long-term future. Perhaps you know that you have a fairly stable level of happiness that you tend to bounce back to and that it's unlikely that even a meaningful career shift will increase or decrease your expected levels of life satisfaction in the long-term. So you might think the most likely states you could be in are ones in which A is temporarily worse than B or B is temporarily worse than A, before they return to some common baseline.



[19] Some option-preserving interventions might have been discovered and implemented at a later date. If so, the value of implementing them now rather than later is the difference in value created between those two times. If an intervention would avert a single existential disaster that would not have been averted otherwise and cannot be implemented later, however, the value of implementing it now rather than later is the difference between the value of the entire future and nothing. Therefore the latter type of intervention is much more valuable than the former on a longtermist view.

[20] This idea is discussed, among others, by Moore (1903: sec. 93) and Smart (1973: 33).

Even if the states in the bad decision world take longer than one time step to get better, the expected values of choosing A and B will be more and more similar as time goes on, despite the fact that your choice could make a significant causal difference in the short run. In this example, longtermists and neartermists would agree that the choice of A over B should be driven by near-term considerations, since both choices result in a similar long-term future.

If causal diffusion is true in a decision situation, longtermist and neartermist decision makers will largely agree on what we should do. Tarsney (2022) calls this the *control challenge* to longtermism. Roughly speaking, the worry is that we just cannot control the long-term future because we cannot act on it in ways that are lasting and positive. Therefore, even if we do not intrinsically discount the long-term future, we should focus on what we can control and select actions based on their near-term consequences. So the claim of causal diffusion can justify myopia for longtermists.

But is there widespread causal diffusion in the decision situations we actually face? There are reasons to be skeptical. As Greaves (2016: 314–315) points out, our actions often affect the identities of future people. Every small decision we make can affect, in many ways, which particular children will be conceived, which in turn changes the population of people that will exist in the future. These future people will make different choices themselves, resulting in diverging futures which might be radically different in other respects.[21] This argument suggests that the causal consequences of our actions often do not wash out, which weakens the case from causal diffusion for myopia.

Tarsney (2022: 10) also points out that if there were widespread causal diffusion, we should expect it to be hard or impossible to improve the long-term future even if we can see the objective likelihoods of what will happen far into the future given all the different actions available to us. But it seems implausible that this kind of knowledge about the future would be of such little help to us. Surely, if we knew all the objective likelihoods of everything that will happen as a result of our actions, we could identify actions which make a lasting difference. For example, it seems we could make a lasting difference by increasing the likelihood that a particular person comes into existence or by preventing incidents that lead to a global war.

Therefore, although the claim that the causal consequences of our actions wash out over time would certainly justify myopic decision-making on the part of longtermist agents, there are reasons to be skeptical whether causal diffusion accurately characterizes the decision situations we actually face. There are reasons to think that at least some of our actions can have important consequences that do not wash out over time.

4.2 Epistemic diffusion as a reason for myopia

Even if we accept that some of our actions can have long-lasting effects, it seems exceedingly hard to predict what these effects are. We might agree with Greaves that our actions can have long-lasting effects by affecting the identities of future people, but it seems almost

[21] Greaves (2016: 314–315) writes that 'even our most trivial actions are very likely to have unforeseen *identity-affecting effects* [...] But once my trivial decision has affected *that*, it equally counts as causally responsible for *everything the child in question does during his or her life* [...] and of all the causal consequences of all *those* things, stretching down as they do through the millennia.'

impossible to predict how our actions will affect who comes into existence and what those people will do differently from those that would have existed were it not for our actions.

Epistemic diffusion is the claim that even if the causal effects of each action might not decay over time, it becomes harder and harder to predict the effects of any given action with increasing temporal distance.[22] Consider again your decision about whether to apply for academic jobs or pursue a promising position outside academia. You are reasonably confident about the causal effects of your choice on the immediate future. You are a bit less confident about what will happen as a result of your choice in one year. You are still less confident about what will happen in two years, and so on. As possibilities branch out further and further, it quickly becomes virtually impossible to predict which action will result in a better outcome. We can imagine that you really care about the far-future consequences of your decision, but there's simply no way for you to predict what these consequences will look like. You are *clueless* about the long-term effects of your actions (Lenman 2000; Greaves 2016).

How hard is it to predict the future in the real world? Results from physics and chaos theory suggest that the answer is: very hard. Many physical systems in the real world are chaotic, which means that slightly different initial conditions will lead to large differences in behavior over time. Kravtsov (1993: 195–196) gives the example of idealized colliding gas particles, whose direction becomes virtually unpredictable after only a few collisions as small initial uncertainties are magnified. Werndl (2009) argues that in chaotic systems, past events are approximately probabilistically irrelevant for predicting sufficiently far-future events. This means that if we consider events which are far enough in the future, no amount of information about the past makes any difference to the probability of these future events.

To flesh out this picture, imagine we carved the future into time slices of equal length t_0, $t_1, \ldots t_n$. Our information about what the world will look like decreases as the index of these time slices increases. So, for any n, the likelihood that a given future time slice t_j is of value n conditional on undertaking action A may converge on the likelihood that t_j is of value n conditional on undertaking action B as j increases. If this is the case then the expected value difference between the outcomes of action A and action B goes down as the temporal index increases. Moreover, there may be some future time t_i after which the expected value of each future time slice is *equal* conditional on undertaking action A and action B. Therefore the expected value of action A and action B will be entirely determined by events that occur before time t_i.[23]

Epistemic diffusion is different from causal diffusion. You might *know* that your choice will lead to lasting causal differences, for example by affecting the identities of future people. However, you might have no way of predicting what these lasting causal differences will be. To understand this point, it is again useful to consider chaotic systems. In such systems, small differences in initial conditions can make large causal differences over time, so there is no causal diffusion. However, we have no good way of predicting what the future behavior of a chaotic system will look like even if we know its past behavior, so there is epistemic diffusion. Again, we can make this precise by drawing on Werndl (2009) and her

[22] Thorstad (2021: 10–13) discusses a similar argument ('the washing out argument') and arguments which suggest that many of our actions do not make a persistent impact on the future. Tarsney (2022) calls this 'epistemic challenge' to longtermism.

[23] For moral theories that care about features of the world not reflected in their value, we could replace the expected value of the world at each time slice with the expected way the world is at that time slice.

point that in chaotic systems, information about the past is approximately probabilistically irrelevant about events in the far future.

Epistemic diffusion reduces the value of trying to take actions that affect the long-term future. We can know the near-term consequences of actions and compare them with some accuracy, but if we think that the long-term consequences of our actions are mostly influenced by causal interactions that we cannot predict and that are independent of our actions, we may end up thinking there is no reason to assign higher value to one long-term future than the other, and therefore assign the two possible outcomes similar value beyond a certain point in time.[24] Given this, epistemic diffusion is a reason for longtermists to behave myopically.

The longtermist might respond that if we are clueless about the long-term consequences of our actions, we have a strong incentive to seek out information which resolves our uncertainty about the future. This incentive is stronger if one accepts longtermism than if one accepts neartermism, and so cluelessness does not cause longtermists to behave more myopically.

There is definitely some truth to this response. Suppose the longtermist could invest resources into the task of building a perfectly reliable oracle which could peer into the future. Since the longtermist cares about the long-term consequences of her actions, she should be willing to invest a large amount of resources into the construction of such an oracle if the probability it can be achieved is high enough. In contrast, a decision maker who intrinsically discounts the future would be willing to invest fewer resources into the construction of the oracle even if the likelihood of success were held fixed.[25]

However, epistemic diffusion means not only that we are uncertain about the future, but also that the information we can learn about the future is noisy and unreliable and generally gets *more* noisy and unreliable the farther we try to peer into the future. Noisy and unreliable information has less value since it is less useful as a guide to making good decisions.[26] As explained above, we can make this precise by saying that in chaotic systems, information about the past is approximately probabilistically irrelevant for events in the far future. If the information is irrelevant to the future, the value of learning the information is zero, since there is no way the information could make a difference to our decision.

More noisy information has less value insofar as it can be expected to move our credences less far from our priors. If the information we can gather about the long-term future is increasingly noisy and unreliable, the value of that information will decrease rapidly relative to its cost. So in the presence of epistemic diffusion, placing higher value on information about the future may cause longtermists to behave only marginally differently from neartermists. If the cost per bit of information about a time t_j increases exponentially with j, for example, then it is likely that the longtermist will cease investing in information about the future not long after the neartermist similarly ceases to invest in this information.

[24] This is similar to what Greaves (2016) calls the problem of *simple cluelessness*.

[25] Incidentally, this seems to imply that the construction of a time machine would also be a much higher priority for longtermist decision makers relative to neartermist decision makers.

[26] To illustrate how noise decreases the value of information, consider again the case in which you are uncertain whether $100 is in a box on your left or your right. You can learn some noisy information which increases your credence in the left or right box by 1% in expectation. Since the expected improvement in your guess is small, you should not pay more than $2 in order to learn this information. In the limit where the information is probabilistically independent of where the money is, the value of learning the information is zero.

Epistemic diffusion will cause longtermist and neartermist decision makers to be in greater agreement about what to do. We can say all this while preserving the intuition that if it were somehow feasible to peer into the distant future and learn which action has the best long-term consequences, we would be required to do so.[27] The point is that this is not, in fact, feasible.

It might be objected that cluelessness is only a problem for those that reject precise Bayesianism. Greaves (2016) writes that a decision maker is clueless if she has no idea which of the available actions has the best overall consequences and claims that it is hard to even formulate this problem for orthodox Bayesian models:

> It is not at all obvious on reflection, however, what the phenomenon of cluelessness really amounts to. In particular, it (at least at first sight) seems difficult to capture within an orthodox Bayesian model, according to which any given rational agent simply settles on some particular precise credence function, and the subjective betterness facts follow. (Greaves 2016: 336)

The model we have assumed is an orthodox Bayesian model. We have a subjective probability distribution which settles, at each time step, the expected values of all available actions. So in a sense, our agent can never be clueless: for any pair of actions, her subjective probabilities settle which action has the higher expected value.

In contrast, our discussion shows that even those working within an orthodox Bayesian model face a version of the problem of cluelessness. Epistemic diffusion means that the expected value of our actions becomes hard to distinguish beyond a certain point in time. Since information about the long-term consequences of our actions is noisy and unreliable, gathering information about their long-term consequences is often prohibitively costly.[28] If the world is a highly chaotic system, we can say something even stronger: information about the past is approximately probabilistically irrelevant to the long-term future, and so the value of information about the long-term future goes to zero.[29] This means that our decisions must be determined more by the predictable near-term consequences of our actions than by their long-term consequences. Therefore epistemic diffusion pushes longtermists towards more myopic behavior.

4.3 Moral uncertainty as a reason for myopia

Whether we should discount the future and by how much is a difficult moral question that longtermists and neartermists disagree on. Plausibly, we should not be perfectly confident that we know the right answer to this question. It is probably rational to have at least some

[27] After all, if an omniscient oracle gave us a scroll that detailed exactly what the consequences of an action would be, surely we would read the scroll for as long as is feasible before making our decision, rather than skimming the first few lines and concluding that we know enough to decide how to act.

[28] There are other aspects of cluelessness which are not captured by our orthodox Bayesian model. For example, we might be ignorant of what the space of possibilities looks like. This kind of ignorance is discussed in the literature on unawareness (Steele and Stefánsson 2021) and might be part of what Greaves (2016: 323) calls 'complex cluelessness'.

[29] Economists also recognize that myopic behavior reflects not only intrinsic time preference but also uncertainty about the future (Yaari 1965; Sozou 1998; Gabaix and Laibson 2022).

moral uncertainty about the correct pure discount rate to have. How exactly to incorporate this kind of uncertainty into our model is a non-trivial question.[30]

One way to incorporate moral uncertainty about temporal discount rates would be to have a probability distribution over possible discount rates that represents our confidence that each is correct and use the 'expected discount rate' with respect to this distribution for the purpose of decision-making. In this simple model, moral uncertainty about the correct discount rate will almost certainly push us towards more myopic behavior.

There are many sources of support—both testimonial and theoretical—for positive future discount rates but almost no sources of support for *negative* future discount rates, i.e. discount rates that imply that we should value a unit of value tomorrow more than a unit of value today.[31] So even if we find the moral arguments for not discounting the future quite convincing, we will likely assign a small positive probability to discounting the future and thus end up with a small-but-positive expected discount rate for the purpose of decision-making. Over a sufficiently long time horizon, even a small-but-positive effective discount rate could become quite significant.

There are other ways in which we can incorporate moral uncertainty about discounting rates into our decision-making. One alternative would be to say that we are morally uncertain about different moral theories: theories that have associated discount rates. Each moral theory M assigns a value $V_m(A)$ to each action A and we value each action by combining these weighted values according to the credence we assign to the different moral theories.[32]

If we use the procedure above, we will end up with less discounting than if we use expected discounting rates. Actions whose effects are only felt far in the future will be assigned little value if we employ an expected discount rate. But since the longtermist would assign high value to these actions, using expectations over moral theories means that they will continue to be of importance. However, we still end up with greater myopia in response to moral uncertainty.

Alternatively, one could adopt a view on which everyone behaves in accordance with the discount rate they find most plausible.[33] This would mean that many people would act in accordance with neartermist discount rates but a small subset would act in accordance with longtermist discount rates, rather than everyone adopting an uncertainty-weighted discount rate.[34] In this case, as a group, we would still end up with a considerable amount of myopia but those sufficiently confident in longtermism would not behave more myopically.

There are also other ways in which moral uncertainty might push us towards myopia. We are morally uncertain about lots of things: axiological questions like what the correct theory of population ethics looks like (Greaves and Ord 2017), deontic questions like whether we should always maximize expected value or whether we ought to respect rights and duties and so on. These kinds of moral uncertainty seem to weaken the longtermist's

[30] Lockhart (2000) and MacAskill, Bykvist, and Ord (2020) discuss many aspects of moral uncertainty.

[31] Mogensen (2022b) discusses an argument for a positive discount rate and Cowen and Parfit (1992) argue against discounting.

[32] Thanks to an anonymous reviewer for this suggestion.

[33] This is reminiscent of the 'My Favorite Theory' approach to moral uncertainty, where one simply follows the verdicts of the moral theory one finds most plausible (Lockhart 2000: 42). While this approach faces objections, it has also been defended (Gustafsson and Torpman 2014).

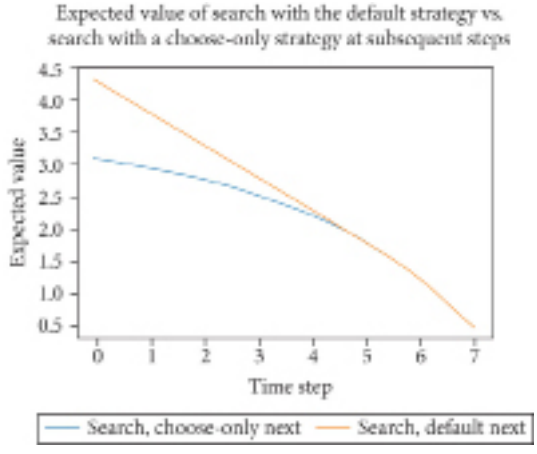
[34] Not splitting the difference could be optimal from a group rationality point of view, since (i) it leads to a wider portfolio of actions, and (ii) agents can see the outcomes of the actions of other agents and update on these. If there is a lot of disagreement about what the outcome of an experiment will be, it may be better to have groups defend the view they think is most likely rather than splitting the difference, so that their evaluation of the independent evidence (i.e. of non-testimonial evidence) can be evaluated by onlookers.

case that we should mostly focus on improving the long-term future. As an illustration, we might assign some credence to the view that we should pay special attention to the rights of existing people when making decisions or to person-affecting views in population ethics.[35] Once we take this uncertainty into account, we will plausibly end up with more myopic decisions, although different ways to incorporate moral uncertainty will, as we saw above, lead to different degrees of myopia.

4.4 The optimism-pessimism dilemma

Longtermists prioritize actions that cause a persistent improvement to the future. But the expected value of actions depends on the behavior we expect from future agents, including ourselves. If we know that someone else is going to destroy the world in 10 minutes with 100% certainty, then the value of saving the world now is just the value of those 10 minutes.

We can illustrate this using the trap game from subsection 3.2. When we first introduced this game, we calculated the expected value of searching for traps with the assumption that the agent would continue to use a policy that maximized undiscounted future value for the rest of the game. Now let's consider the value of searching for traps if the agent expects to switch to a 'choose-only' strategy after the current time step, i.e. a strategy where she will always choose left or right rather than searching for traps.



The expected value of searching early in the game is lower if she expects to adopt a choose-only strategy after the current time step. This is because a choose-only strategy has lower expected future value, given the greater risk of stepping on a trap and ending the game.

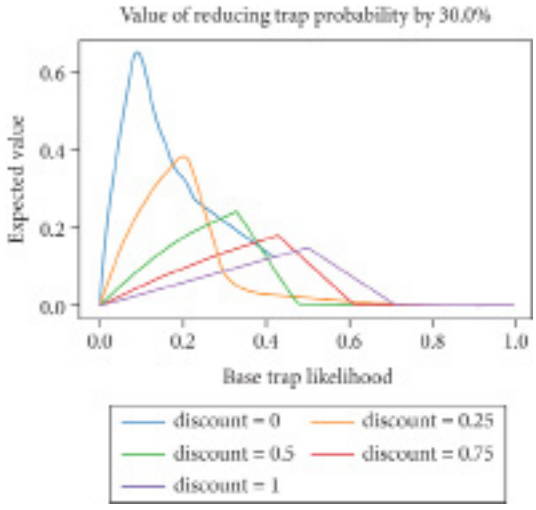
This leads to a dilemma for longtermism discussed by Thorstad (2022). On the one hand, suppose we are 'optimists' about the future. We believe that the likelihood of existential catastrophes or harmful value lock-in is low. This could be because we have trust in future people—we think they will have the will and the ability to avert future catastrophes—or

[35] Frick (2017) and Ord (2020: 259–264) discuss objections to longtermism from person-affecting views in population ethics.

because we simply think that by default these outcomes are unlikely.[36] This seems like good news from the longtermists' perspective, but it also reduces the expected value of taking actions with an eye to the long-term future, since it is unlikely that our actions can improve the future much. Therefore optimism about the future will increase the relative value of behaving myopically from a longtermist perspective.

On the other hand, suppose we are 'pessimists' about the future. We believe that the likelihood of existential catastrophes or harmful value lock-in is high. This could be because we lack trust in future people or because we believe these outcomes are likely by default. This seems terrible from the longtermist perspective. But it also reduces the expected value of taking actions with an eye to the long-term future, since it is unlikely that our actions can save such a future from likely disaster (Thorstad 2022). Therefore pessimism about the future will increase the relative value of behaving myopically from a longtermist perspective.

We can illustrate this with another variant of the corridor game. Suppose an agent can take an action before the start of the game that will reduce the likelihood of traps that she will face in the game by 30%. Now suppose that the starting rate of traps can be anywhere from 1 (guarantee of a trap on the next time step) to 0 (no chance of a trap on the next time step). Below we show the expected value of her taking this risk-reducing action before the start of the game.



The expected value of reducing the trap probability by 30% depends on the base likelihood of traps. In 'high trap' worlds where the risk of traps is close to 1, the value of reducing traps is zero because this reduction doesn't make enough of a difference to the actions available to the agent (they are likely to adopt a search-heavy strategy regardless). In 'low trap' worlds where the risk of traps is already close to zero, the value of reducing traps is also lower, because this also doesn't make much of a difference to the actions available to the agent (they are likely to adopt a choose-heavy strategy regardless).

The act of reducing trap probabilities is most valuable in cases where base trap likelihood is somewhere in between these two extremes. In other words, it is most valuable in situations that are neither fully pessimistic nor fully optimistic. The value of reducing the trap probability

[36] For example, Aschenbrenner (2020) discusses a model of growth according to which existential risk reduces to zero in the long-term, after initially growing.

in these intermediate cases also increases as the agent's discount rate decreases, i.e. the more longtermist the agent is, the more valuable the action is in these intermediate cases. So if you are an agent who does not discount the future, both optimism and pessimism will push you towards myopic behavior. Longtermism will only result in hyperopia if it is combined with an intermediate view that is neither overly optimistic nor overly pessimistic about the future.[37]

5 Conclusion

We began this chapter by noting that longtermists face a potential dilemma: if longtermists are too myopic—if they say that we should give the same weight to present improvements relative to the long-term future as neartermists—then longtermism faces an objection from irrelevance. But if longtermists are too hyperopic—if they say that we should be willing to neglect present people in favor of improving the lives of far-future people—then longtermism faces an objection from future fanaticism.

We have explained some of the main considerations that will push longtermists towards hyperopia and myopia. Because they place more weight on the far future, longtermists will prioritize getting information about the future and preserving option value more than neartermist agents. This causes longtermists to place greater importance on things like mitigating existential risks. However, we have argued that causal diffusion, epistemic diffusion, moral uncertainty, and optimism or pessimism will all push longtermist agents to behave more myopically.

Given the strength of the reasons in favor of myopia, we believe that longtermism presents a less radical deviation from common moral intuitions than one might initially expect. Whether these considerations are so strong that they raise the specter of irrelevance, so weak that they cannot banish the specter of future fanaticism, or just strong enough to help the longtermist avoid the worst versions of both, is left to the reader to decide.[38]

References

Aschenbrenner, L. (2020), 'Existential Risk and Growth', GPI Working Paper No. 6-2020 (Global Priorities Institute, Oxford University).

Beckstead, N. (2013), 'On the Overwhelming Importance of Shaping the Far Future', PhD thesis, Rutgers University.

Beckstead, N. and Thomas, T. (2021), 'A Paradox for Tiny Probabilities and Enormous Values', GPI Working Paper No. 7-2021 (Global Priorities Institute, Oxford University).

Bertsekas, D. P. (2005), *Dynamic Programming and Optimal Control*, Volume 1, 3rd edition (Athena Scientific).

Buchak, L. (2010), 'Instrumental Rationality, Epistemic Rationality, and Evidence-Gathering', in *Philosophical Perspectives* 24/1: 85–120.

Cowen, T. and Parfit, D. (1992), 'Against the Social Discount Rate', in P. Laslett and J. Fishkin (eds.), *Philosophy, Politics, and Society*, sixth series (Yale University Press), 144–161.

Frick, J. (2017), 'On the Survival of Humanity', in *Canadian Journal of Philosophy* 47 /2–3: 344–367.

[37] If we live in a hinge point in history—an especially influential time for the future of humanity—then having expectations at this sweet spot in between optimism and pessimism might be rational. This idea has been discussed by Parfit (2011: 616), MacAskill (2020), Karnofsky (2021), and Mogensen (2022a).

[38] Equal contributions from both authors. The authors would like to thank Kamal Ndousse for reviewing the content and code in this chapter and offering helpful suggestions, and an anonymous reviewer for their helpful comments and suggestions.

Gabaix, X. and Laibson, D. (2022), 'Myopia and Discounting', NBER Working Paper No. 23254 (National Bureau of Economic Research).
Greaves, H. (2016), 'Cluelessness', in *Proceedings of the Aristotelian Society* 116/3: 311–339.
Greaves, H. and MacAskill, W. (2021), 'The Case for Strong Longtermism', Global Priorities Institute Working Paper 5-2021 (Global Priorities Institute, Oxford University).
Greaves, H. and Ord, T. (2017), 'Moral Uncertainty About Population Axiology', in *Journal of Ethics and Social Philosophy* 12/2: 135–167.
Good, I. J. (1967), 'On the Principle of Total Evidence', in *British Journal for the Philosophy of Science* 17/4: 319–321.
Gustafsson, J. E. and Torpman, O. (2014), 'In Defence of My Favourite Theory'. *Pacific Philosophical Quarterly* 95 (2): 159–174.
Kadane, J. B., Schervish, M., and Seidenfeld, T. (2008), 'Is Ignorance Bliss?', in *Journal of Philosophy* 105/1: 5–36.
Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998), 'Planning and Acting in Partially Observable Stochastic Domains', in *Artificial Intelligence* 101/1–2: 99–134.
Karnofsky, H. (2021), 'All Possible Views About Humanity's Future Are Wild', *Cold Takes,* https://www.cold-takes.com/all-possible-views-about-humanitys-future-are-wild/ (accessed January 31, 2025).
Kravtsov, Y. A. (1993), 'Fundamental and Practical Limits of Predictability', in Y. A. Kravtsov (ed.), *Limits of Predictability* (Springer), 173–203.
Lenman, J. (2000), 'Consequentialism and Cluelessness', in *Philosophy & Public Affairs* 29: 342–370.
Lockhart, T. (2000), *Moral Uncertainty and Its Consequences* (Oxford University Press).
MacAskill, W. (2020), 'Are We Living at the Hinge of History?', GPI Working Paper No. 12-2020 (Global Priorities Institute, Oxford University).
MacAskill, W. (2022), *What We Owe the Future* (Basic Books).
MacAskill, W., Bykvist, K., and Ord, T. (eds.) (2020), *Moral Uncertainty* (Oxford University Press).
Madani, O., Hanks, S., and Condon, A. (2003), 'On the Undecidability of Probabilistic Planning and Related Stochastic Optimization Problems', in *Artificial Intelligence* 147/1–2: 5–34.
Mahtani, A. (2019), 'Imprecise Probabilities', in R. Pettigrew and J. Weisberg (eds.), *The Open Handbook of Formal Epistemology* (PhilPapers Foundation), 107–130.
Mogensen, A. (2022a), 'The Hinge of History Hypothesis: Reply to MacAskill', GPI Working Paper No. 9-2022 (Global Priorities Institute, Oxford University).
Mogensen, A. (2022b), 'The Only Ethical Argument for Positive δ? Partiality and Pure Time Preference', in *Philosophical Studies* 179/9: 2731–2750.
Monton, B. (2019), 'How to Avoid Maximizing Expected Utility', in *Philosophers' Imprint* 19/18: 1–25.
Moore, G. E. (1903), *Principia Ethica* (Cambridge University Press).
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Parfit, D. (1984), *Reasons and Persons* (Oxford University Press).
Parfit, D. (2011), *On What Matters: Volume Two* (Oxford University Press).
Puterman, M. L. (1994), *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (Wiley).
Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory* (Graduate School of Business Administration, Harvard University).
Russell, S. and Norvig, P. (2018), *Artificial Intelligence: A Modern Approach* (Pearson).
Savage, L. J. (1954), *The Foundations of Statistics* (Wiley Publications in Statistics).
Smart, J. J. C. (1973), 'An Outline of a System of Utilitarian Ethics', in J. J. C. Smart and B. Williams, *Utilitarianism: For and Against* (Cambridge University Press), 3–74.
Sozou, P. D. (1998), 'On Hyperbolic Discounting and Uncertain Hazard Rates', in *Proceedings of the Royal Society B: Biological Sciences* 265/1409: 2015–2020.
Steele, K. and Stefánsson, H. O. (2021), *Beyond Uncertainty: Reasoning with Unknown Possibilities* (Cambridge University Press).
Tarsney, C. (2022), 'The Epistemic Challenge to Longtermism', GPI Working Paper No. 3-2022 (Global Priorities Institute, Oxford University).
Thorstad, D. (2021), 'The Scope of Longtermism', GPI Working Paper No. 6-2021 (Global Priorities Institute, Oxford University).
Thorstad, D. (2022), 'Existential Risk Pessimism and the Time of Perils', GPI Working Paper No. 1-2022 (Global Priorities Institute, Oxford University).
Titelbaum, M. (2019), 'Precise Credences', in R. Pettigrew and J. Weisberg (eds.), *The Open Handbook of Formal Epistemology* (PhilPaper Foundation), 1–55.
Werndl, C. (2009), 'What Are the New Implications of Chaos for Unpredictability?', in *British Journal for the Philosophy of Science* 60/1:195–220.
Wilkinson, H. (2022), 'In Defense of Fanaticism', in *Ethics* 132/2: 445–477.
Yaari, M. E. (1965), 'Uncertain Lifetime, Life Insurance, and the Theory of the Consumer', in *The Review of Economic Studies* 32/2: 137–150.

18

Minimal and Expansive Longtermism

Hilary Greaves and Christian Tarsney

1 Introduction

Strong longtermism (hereafter, simply *longtermism*) is, roughly, the thesis that 'impact on the far future is the most important feature of our actions today' (see Beckstead 2013: 1–3; Beckstead 2019: 80; Greaves and MacAskill this volume).* This chapter examines the question of how expansive the most plausible form of longtermism is. Roughly for now, the issue we see is that (i) the standard argument for longtermism implies only a rather minimal form of the thesis, (ii) many authors and real-world agents subscribe to a far more expansive version of longtermism, (iii) the difference between the two has not previously been in sharp focus, and (iv) for various reasons, the difference between the minimal and more expansive versions of longtermism matters. The aims of our essay are to highlight the differences between 'minimal' and more 'expansive' forms of longtermism, and to conduct some exploration of what arguments for the more expansive forms might look like (without either endorsing or rejecting those arguments).

Before explaining this distinction between minimal and expansive longtermism, let us rehearse the 'standard argument' for longtermism alluded to above.[1] In this standard argument, the motivation for longtermism comes from considering what we will call *technological existential risks* (hereafter, *technological x-risks*). An *existential risk* (or *x-risk*) is a risk of an *existential catastrophe*, that is, either premature human extinction or another irreversible outcome that is similarly bad to extinction.[2] A *technological x-risk* is an existential risk resulting from advanced (present or future) technology.

The standard argument for longtermism then runs as follows.

First, technological x-risks are at worryingly high levels. Areas of concern include, for example, nuclear weapons, artificial intelligence, and biotechnology. For the first time in history, we are creating technologies that could destroy humanity's entire future (see Rees 2003; Posner 2004; Häggström 2016; Russell 2019; Ord 2020).

* The quote is from Greaves and MacAskill (this volume). Beckstead's thesis is that 'what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, or trillions of years' (2013: 1; 2019: 80).

[1] For reasons of space, our presentation of the argument is only rough. For more careful presentations, see Beckstead (2013; 2019) and Greaves and MacAskill (this volume). Our aim is to consider what form of longtermism is most plausible *given that* this standard argument is correct. Thus, we will take the correctness of its premises for granted throughout the chapter. This is not to claim that we ourselves are certain of their correctness, or to deny that they face reasonable objections.

[2] Existential catastrophe is often defined as 'the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development'. For alternative formulations, see Cotton-Barratt and Ord (2015: 1–3), Ord (2020: 37), Greaves (2024). In line with the standard argument, we assume that human extinction would be very bad (at least in expectation). For a contrary view, see for instance, Arrhenius and Bykvist (1995: ch. 3) Althaus and Gloor (2019), and Benatar (2006: 183–200).

Hilary Greaves and Christian Tarsney, *Minimal and Expansive Longtermism* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Hilary Greaves and Christian Tarsney 2025.
DOI: 10.1093/9780191979972.003.0018

Second, there are things that society could do to significantly reduce these risks. For example, we could scale up efforts to avoid great power conflict, development of technical safety tools for artificial intelligence, pandemic preparedness, and legal regulation of dangerous technologies. These opportunities correspond to extremely high-value options that are available to various agents in certain decision situations. For example:

- Philanthropists with no constraints on what causes they support could provide funding to scale up these efforts.
- Talented individuals could devote their careers to these efforts.
- Policymakers could direct more public funding toward these efforts, or implement the appropriate regulations.

Third, the stakes are enormous. The risks in question threaten not just the present generation, but the *whole of the future of humanity*—a future that, if things go well, could continue for millions or billions of years. So anything we can do to reduce x-risk by even a tiny amount will carry enormous expected value—which is the metric by which orthodox decision theory evaluates options under uncertainty.[3]

Fourth, these enormous stakes lie mainly in the far future (more than 100 or even 1,000 years hence). A typical lower-end estimate for the expected number of future lives is 10^{16}. Of these, only around 10^{10} occur in the next 100 years. So the next century accounts for only around 0.0001% (or less) of the value that would be destroyed by an existential catastrophe.

Finally: Crunching the numbers, we find that the far-future expected benefits of mitigating technological x-risks are vastly greater than any plausible expected benefit one could deliver within (say) the next 100 years (what we will call the 'neartermist benchmark').[4] Some other type of action might be even better than mitigating technological x-risk, but if so, it would have to share the feature that most of its (expected) benefits are in the far future. Thus, impact on the far future is the most important feature of the most important actions that agents today can take. The far future is, in this sense, the most important feature of 'our actions today'—as per longtermism.[5]

Various premises and moves in this argument might be disputed, but that is not our concern here. The motivation for the present essay arises rather from the fact that the final step in the argument, even if correct, at least carries some risk of being importantly misleading.

[3] In line with the standard argument for longtermism, we assume the correctness of expected value maximization for concreteness and simplicity. Thus, when we talk about the *value*, *impact*, *benefits*, etc. of an option, we always mean *expected* value/impact/benefits, unless otherwise specified. Other approaches to *ex ante* evaluation might or might not yield similar conclusions (see for instance Pettigrew 2024, Tarsney forthcoming). But investigating the extent to which the present discussion is affected by choice points within decision theory lies beyond the scope of this essay.

[4] This 'neartermist benchmark' might correspond, for instance, to the effects of bednet distribution on saving the lives of already existing people. For further discussion, see Greaves and MacAskill (this volume).

[5] All the longtermist theses we consider in this chapter ('minimal longtermism', 'expansive longtermism', and '(strong) longtermism' *simpliciter*) should be understood as *ex ante* axiological theses, i.e., theses about the value of possible actions relative to the agent's evidence. Thus, the preceding argument (i) does not claim to establish conclusions about what agents *ought* to do, or about moral permissibility, or to deny that various non-consequentialist considerations may bear on these further questions; and (ii) does not depend on the claim that we *can in fact* make the difference between existential catastrophe occurring vs. not.

What is most directly supported by the above line of argument is what we will call *minimal longtermism*:

> *Minimal longtermism*: It is extremely important, for reasons concerning the far future, that key decision-makers (namely: philanthropists, talented individuals choosing careers, and some policymakers) take actions to direct significantly more resources toward addressing the technological x-risks mentioned above.

Minimal longtermism is 'minimal' in at least three respects.

First, it may recommend only one narrow type of action (or 'intervention'): interventions that aim quite directly at reducing technological x-risks (such as funding to scale up technical artificial intelligence (AI) safety work). The examples to which the standard argument appeals are exclusively of this type. It is consistent with that argument, and with minimal longtermism, that the only way of significantly improving the far future (in expectation) is to mitigate technological x-risks.

Second, minimal longtermism may concern only a narrow class of choice situations. The standard argument considers only the decision situations of (i) philanthropists or policymakers deciding how to allocate resources between different areas (and specifically, between x-risk mitigation and other, potentially more 'neartermist' objectives), (ii) individuals choosing a career path, and (iii) policymakers enacting regulation in a few very specific areas (e.g., AI and biotechnology). While these decision situations may be especially important, they do not, by any natural count, encompass a majority of the decisions that present-day agents face. It is consistent with the standard argument, and with minimal longtermism, that in most decisions the far future is not a relevant issue.

Third, minimal longtermism may call for only a relatively modest reallocation of resources: for redirecting, say, less than 2% of world GDP toward addressing technological x-risks. The standard argument estimates the cost-effectiveness of opportunities to further mitigate technological x-risks at or near the current margin, but says nothing about how fast this marginal cost-effectiveness would decline if far more resources were allocated in that direction. It is consistent with the standard argument, and with minimal longtermism, that if (say) 2% of world GDP were spent on carefully chosen initiatives to mitigate technological x-risks, it would no longer be the case that attainable far-future benefits exceed the near-future benefits that are attainable for the same cost.

We said that the standard argument *establishes* only minimal longtermism. But some authors and organizations either explicitly endorse, or appear from their actions to believe, what we will call *expansive longtermism*. Expansive longtermism goes beyond minimal longtermism on all three of the dimensions just highlighted:[6]

> *Expansive longtermism*:
>
> 1. There is a very wide variety of ways to greatly improve the expected value of the far future. The full range of interventions with this property is many and varied, united by little or

[6] One could, of course, side with the 'expansive' over the 'minimal' view on some but not all of these three dimensions.

nothing except that they all significantly improve the far future in expectation. Actions that mitigate technological x-risks are just particularly clear-cut examples.

2. Relatedly, possible effects on the far future are the main determinant of expected value comparisons in *nearly all* decision situations faced by *nearly all* present-day human agents.
3. All this is not only true at the current margin, but is likely to remain true for at least the next several decades (barring an existential catastrophe), even if massively more resources (say, more than 50% of world GDP) come to be directed in ways that are optimized for the far future. The truth of longtermism arises from quite fundamental features of our situation in the early 21st century, rather than merely from an easily remediable failure to adequately fund and implement a few key safety measures related to new technologies.

Our view is that, while expansive longtermism *may* be true, the arguments for expansive longtermism are significantly less robust and significantly more speculative than the arguments for minimal longtermism. This is the thought that motivates the present essay. The essay's intended contributions are twofold. First, we call attention to the important differences between minimal and expansive longtermism, and to the fact that the standard argument establishes only the former. Second, we explore potential arguments for versions of longtermism that are more expansive than the minimal version, in any or all of the three ways just outlined. However, for the most part, we do not defend firm conclusions regarding which of those arguments ultimately hold water (a matter on which the two authors of this essay often incline in somewhat different directions).

The gap between minimal and expansive longtermism is important for at least three reasons. First, the question of *how wide a range of interventions* can significantly improve the far future in expectation relates to the question of whether longtermism is a helpful, or instead a misleading, organizing concept. If the only ways of significantly improving the far future in expectation are targeted actions to mitigate technological x-risks, then even if it is *true* that impact on the far future is the most important feature of our actions today, it would be equally true and more informative to say that impact on technological x-risks is the most important feature of our actions today.[7] Currently, however, at least some philanthropic organizations explicitly adopt the more expansive 'longtermist' framing.[8] This risks over-rating a large collection of hopelessly intractable would-be 'longtermist' projects, by lumping them together with the extremely valuable project of technological x-risk mitigation.

Second, the remaining two dimensions on which expansive longtermism differs from its minimal cousin speak to an important source of unease about the longtermist worldview: the thought that longtermism threatens to be radically revisionary of both public and

[7] The idea that 'existential risk mitigation' is a more useful focus or organizing concept than 'longtermism' (even if longtermism is true) comes up frequently in non-academic discussions of longtermism. See for instance Alexander (2022), Nanda (2022), and Yglesias (2022). These discussions often implicitly assume that mitigating technological x-risks is the only or most important way of effectively improving the long-run future.

[8] At the time of writing, for example, the website of Longview Philanthropy states:

> Everything we do is guided by our core values:
>
> 1. **Longtermism:** Our collective future could be extraordinarily good or inordinately bad. We believe this generation has the power to influence which path humanity takes, and it is essential that we act responsibly. (Longview Philanthropy 2021)

private morality on a large scale, to an extent that makes it both epistemically *prima facie* implausible and practically unappealing.

The second dimension (how many decision situations longtermism is true of) affects whether longtermism is best seen as a general moral outlook, or only a truth of much more localized relevance. The former view would be highly revisionary of the way that most people think of their everyday choices. Most of us do not ordinarily think that the far future of humanity bears on our decisions about, for example, what to have for breakfast. Minimal longtermism does not call this assumption into doubt. But on the expansive view, the far future is the primary determinant of the relative values of actions in nearly all decision situations.

The third dimension (*how many resources* would have to be redirected to longtermist causes before longtermism ceases to be true) relates to a concern that longtermism inappropriately deprioritizes the interests of the present generation. If taking longtermist reasoning to its logical conclusion implies that it would be better if (say) 80% of world GDP were directed in ways that are near-optimized for the far future, at massive expense to the present generation, some will regard this as a *reductio ad absurdum* of that longtermist reasoning.

Finally, of course, whether the most plausible version of longtermism is minimal or expansive on the above dimensions is practically important for those who find the standard argument for longtermism compelling. The right strategy for longtermism as an intellectual and social movement might look very different depending on whether its goal is (i) a wholesale moral transformation of society, or merely (ii) a modest redirection of government and philanthropic budgets toward technological x-risk mitigation.

In the next three sections, we explore the three dimensions on which expansive longtermism goes beyond minimal longtermism. With respect to each dimension, we consider what case can be made for the more expansive longtermist thesis, and offer some (tentative, highly uncertain) assessment of that case.

2 How many interventions?

The standard argument for longtermism appeals to back-of-the-envelope calculations of the cost-effectiveness of efforts to mitigate a few key existential risks, primarily in the category of technological x-risks (see Greaves and Macaskill this volume).[9] The case for *these* interventions being extremely cost-effective—and for generating far-future expected benefits that are many times higher than the highest available expected benefits for the near future—seems strong. But if one asks what *else* we can do to predictably improve the far future, *beyond* mitigating these few key risks, it is certainly possible to be unimpressed by what is on offer.

[9] The arguments also sometimes appeal to *natural* x-risks such as those from asteroids and supervolcanoes (see Beckstead 2013: 68–69; Greaves and MacAskill this volume). The cost-effectiveness of mitigating natural x-risks is usually judged to be significantly lower than that of mitigating technological x-risks, because the absolute amount of technological x-risk we currently face is significantly higher than the amount of natural x-risk we currently face (see Ord 2020: 167). But if the space of tractable longtermist interventions is restricted to mitigation of *natural and* technological x-risks, rather than only technological x-risks, the general thrust of our discussion is unaffected.

The problem is that, in general, the project of trying to influence the course of the far future has a strong air of intractability. The further into the future we look, the harder it becomes to predict either what the world will look like in the absence of any intervention on our part, or the effects of any particular present action. Risks of human extinction and other 'existential catastrophes' create an exception to these worries about intractability, since each such risk comes with a strong and clear 'lock-in' mechanism.[10] But most other ways in which we might hope to improve the far future of humanity can be motivated only via significantly more speculative reasoning concerning very long-term causal chains.

That said, we do think that the menu of plausible ways to improve the far future is somewhat longer than *just* highly targeted efforts to mitigate technological x-risks. The remainder of this section explores some possibilities, and gives our own tentative assessments.

2.1 Indirect existential risk mitigation

Even if an impartial concern for the far future recommends no changes to the status quo other than an effective global response to technological x-risks, this effective response might itself comprise a wide variety of interventions. It might include not just 'direct work' (like AI safety research or advocating for nuclear arms reduction) but also various activities that *indirectly* mitigate either particular technological x-risks or technological x-risk in general. Several such indirect strategies have been suggested.

Aschenbrenner (2020) argues that any intervention that speeds up economic growth thereby contributes to mitigating technological x-risk, on the grounds that a growth speedup would reduce the duration of a 'time of perils' during which technological x-risk per unit time is high, and after which it falls to near zero.[11] Aschenbrenner's argument seems inconclusive, however: his analysis considers only one factor, and on balance it seems unclear to us whether faster economic growth would reduce or increase existential risk.[12]

There are other broad social objectives, though, where the directionality (if not the magnitude) of the indirect impact on x-risk seems more clear. For instance, improving the quality of education would produce more informed electorates that might support and demand better policies on issues like climate change, nuclear non-proliferation, and pandemic preparedness. It might also improve the talent pool of future scientists, engineers, bureaucrats, etc., some of whom will work to manage these risks. Likewise, improving collective foresight could help us see existential risks further in advance and in greater resolution—e.g., giving us better predictions of when AGI will be achieved, what kind of actor and what approach to AGI will get there first, and how quickly AGI capabilities will 'take off' thereafter. We could improve foresight on these and other questions by training more professional forecasters and creating better systems to direct the efforts of those forecasters. Finally, general improvements to human moral values and character or to the

[10] For discussion of the preceding points, see Tarsney (2023).

[11] The concept of a time of perils characterized by temporarily elevated levels of existential risk is originally due to Carl Sagan (1994: 173). The claim that we are living in a time of perils is not uncontroversial—for discussion, see MacAskill (2022), Thorstad (2022), and Häggström (this volume).

[12] For example, if the degree of existential risk from AI depends on the values of globally dominant agents at the time when artificial general intelligence (AGI) is first developed (e.g., the governments and large corporations that might participate in its development), those values tend to improve over time, and speeding up economic growth would hasten the advent of AGI, then speeding up economic growth could increase existential risk.

decision-making processes of powerful institutions would presumably reduce technological x-risks by causing future individuals and institutions to act against them more vigorously and effectively.

There are other broad social objectives we might pursue that would indirectly reduce some particular category of technological x-risk. For instance, working to reduce the risk of great power war, while also valuable for many other reasons, would notably reduce existential risks from nuclear weapons. Agents pursuing this goal might aim to reduce regional conflicts among minor powers that might spiral into a great power war; or try to promote and protect democracy in the great powers, since democracies are less likely to go to war with one another (see Chan 1997; Mello 2017); or promote cultural exchange between great powers, to increase public aversion to war.

For another example, Millet and Snyder-Beattie (2017) consider the project of bringing all human and animal health systems up to the minimal standards required by the International Health Regulations, starting from the status quo in which more than half of the world's countries do not have health systems of this standard. One effect of this would be to mitigate existential risks from pandemics, since it would allow rapid detection of and response to a pandemic wherever in the world it originates or spreads. Millet and Snyder-Beattie argue, to our minds convincingly, that even if one counts only the effects of this project on mitigating extinction risks from pandemics, it would save an expected number of lives per dollar that is (on most of the estimates of extinction risk magnitude they survey) significantly better than neartermist benchmarks like the near-term benefits of anti-malarial bednet distributions or direct cash transfers.

A common feature of these indirect x-risk mitigation strategies is that they have substantial near-term cobenefits. Partly for this reason, many of the objectives discussed above already receive a great deal of attention from governments, institutional philanthropists, or other agents. This suggests that it is hard for most agents to make additional progress toward these objectives at the current margin, *perhaps* so hard that—even accounting for their indirect effects on technological x-risks—additional efforts toward these objectives have less marginal benefit than near-term benchmarks. But not all the strategies mentioned above have this character—for instance, improving long-term forecasting and collective foresight is quite neglected, and many health systems in the developing world are dramatically under-resourced. Our view, on the whole, is that there is a highly plausible longtermist case for prioritizing more highly at least some of the objectives discussed above, based on their contribution to technological x-risk mitigation.

2.2 Patient philanthropy

Let us now consider some strategies for improving the far future that are not primarily about reducing existential risk.

One very natural way of improving the far future is to *put more resources into the hands of far-future agents*. This can take two forms. The first strategy, *patient philanthropy*, aims to grow resources over time under compound interest, subsequently putting the proceeds into the hands of future altruistic agents (e.g., philanthropic institutions) for philanthropic projects at that future time. The second strategy, *speeding up growth*, has the less discriminate aim of increasing the wealth of far-future agents generally, most of whom will presumably

use those resources for their own benefit (or to benefit family and others for whom they have partial concern). We will discuss these two strategies in turn.

It has been frequently pointed out that altruistic individuals and philanthropic institutions may be able to do more good, from a temporally impartial point of view, by investing and growing their resources over time, rather than spending any philanthropic resources as soon as they become available (Landesman 1995; Moller 2006; Christiano 2013; Cotton-Barratt 2020). Importantly, this can be the case *even if* the cost-effectiveness of marginal philanthropic projects declines over time as society becomes richer, as the funds in question might grow faster than this marginal cost-effectiveness declines.[13] As emphasized by Trammell (2021a), this should in fact be expected if the majority of other philanthropists operating in the same space have positive rates of time preference. For in that case, the saving rate within the cause area in question will by default be lower than the impartial optimum, so that additional saving at the margin (within the cause area in question) is an improvement from the temporally impartial point of view.

There are various reasonable objections to this idea. For instance, perhaps there will simply be much less acute need in the future than in the present. Long-term philanthropic funds could also be expropriated or otherwise destroyed before they can be put to use. And one must trust that the future agents who manage the disbursement of resources, perhaps centuries from now, will do so wisely and with appropriately altruistic motivations (for a discussion of related worries, see Aird 2020). While these risks are real, however, they are far from certain to occur; while they decrease the amount of long-term philanthropic saving that is optimal, they do not reduce it to zero. It seems very plausible to us that at least at the current margin (with relatively few philanthropic resources invested on very long time horizons), additional long-term philanthropic saving would indeed be an improvement, even when the funds will subsequently be used for (what are then) 'neartermist' projects.[14] If so, this could well be another type of intervention whose far-future expected benefits significantly exceed neartermist benchmarks.[15]

2.3 Accelerating growth

The less discriminate way of transferring resources to the far future is to speed up economic growth.[16] Cowen (2007; 2018) argues that doing this could bring it about that people at every future time are better off than they otherwise would have been, and that since the Earth might support human life and civilization for millions of years to come, this permanent improvement would carry enormous value. Actions that we might take to this end include advocating for pro-growth policies and increases in funding for scientific research and technological development.

[13] There is historical precedent for this. See, e.g., Greaves and MacAskill (this volume).

[14] It might turn out, in the end, that the *most* effective use of a patient philanthropist's funds is to mitigate future existential risks, rather than to provide immediate material benefits to those alive at the time of disbursement (Trammell 2021c: sec. 2.5). In this way, patient philanthropy is *in part* another indirect strategy for existential risk mitigation. But its plausible applications are not limited to existential risk mitigation.

[15] One example of this approach being put into practice is the Patient Philanthropy Fund (see Hoeijmakers 2021).

[16] This could be via any combination of 'level effects' (one-time growth events that increase the *baseline* for future growth, without increasing the future growth *rate*) and 'growth effects' (increases in the growth rate itself over an extended period).

For unoriginal reasons, we are unconvinced by this argument. We will grant the assumption that historically and currently, welfare has generally increased over time as a result of economic growth (with, of course, many local deviations from this general trend). However, it seems far less clear whether we should expect this trend to continue indefinitely into the future. Both common sense and happiness research suggest that beyond a certain point, further economic progress makes little difference to individual welfare (see, e.g., Myers 2000; Kahneman and Deaton 2010). And while *total* welfare at a time can nonetheless increase as long as population increases, there are also limits to the number of people who can sustainably live on Earth at any one time. In that sense, it seems that speeding up economic growth today does not yield exponentially increasing gains over an indefinite future, but only gets us to the inevitable saturation point faster. Even when we aggregate across a long future, the gains to be made from speeding up progress toward a saturation point are limited, and might well be relatively modest, though of course one must ultimately crunch the numbers to be sure (see Beckstead 2013: 67–73; MacAskill 2022: 136–139).[17]

2.4 Space settlement

There may be fairly modest limits to the value of economic growth, and to what we can achieve by speeding up growth, *as long as human civilization remains Earth-bound*. But if humanity eventually begins to settle space, and in particular if it embarks on an indefinite program of interstellar expansion, that is a different matter. First, it makes the upper bounds on accessible resources and sustainable population size astronomically larger. But second, because cosmic expansion means that the number of galaxies we could in principle reach is decreasing every year, the value of those upper bounds depends on how soon we begin to settle space and how fast we expand once we do (see Bostrom 2003). Thus, speeding up space settlement can raise the level at which humanity eventually plateaus, rather than merely getting us to the plateau faster.

Perhaps the strongest longtermist argument for speeding up economic growth is that it speeds up space settlement. But considering the speed of space settlement also suggests other interventions—for example, funding the development of new propulsion technologies and test projects like long-term bases on the Moon or Mars, and lobbying governments to increase funding for their own space agencies.

Bostrom (2003) claims (though with only minimal argument) that it is more important to increase the *probability* that space settlement eventually happens than to increase its *speed*. If this is right, then it primarily supports the conclusion that longtermists should prioritize mitigating risks of premature human extinction and similarly irreversible catastrophes that would prevent humanity from ever settling space. But one could also work to avoid futures in which humanity survives but *chooses* not to settle space, e.g., by supporting positive cultural depictions of space settlement and building legal frameworks that make space settlement positive-sum and attractive to all relevant parties.[18]

[17] In diagrammatic terms, the point is that if total welfare on Earth per unit time will reach a plateau, then speeding up progress is best thought of in terms of shifting an S-curve slightly to the left, rather than shifting the curve of progress slightly upwards for all time. See, e.g., MacAskill (2022: 140). A contrary perspective is suggested by Trammell (2021b).

[18] Even these interventions may count as forms of existential risk mitigation, in the broad sense of 'existential risk' common in the literature. If humanity never settles space, even if it survives happily for hundreds of

However these considerations turn out, the objective of space settlement seems highly likely to expand the menu of interventions whose far-future expected benefits plausibly exceed any benefits attainable in the near future.[19]

2.5 Improving moral values and institutions

Finally (in our incomplete exploration), some longtermists have argued that we can greatly improve the far future in expectation by working to permanently improve either the moral values of human civilization (e.g., via 'moral circle expansion'; see Anthis and Paez 2021) or the quality of our political institutions (see MacAskill 2022: 70–96). Perhaps, for instance, we will eventually reach an equilibrium where one set of moral values is permanently ascendant, but multiple such equilibria are currently possible, and by careful moral reasoning and persuasion, we can positively influence which equilibrium is realized. Or perhaps we are in a similar situation with respect to forms of government. It might be that a world of liberal democracies, a world of totalitarian surveillance states, and a world of extractive quasi-feudal oligarchies are all stable and all currently possible long-term outcomes. In this case, interventions like promoting democracy might have enormous expected long-term benefits, by nudging humanity toward a better long-term equilibrium.

A difficulty for these strategies is that these areas are very crowded: enormous numbers of motivated and talented people have been trying for thousands of years to influence human values and institutions, in hundreds of different, competing directions. There are also no clear mechanisms for long-term persistence of cultural or institutional improvements, as there is for an outcome like extinction. In addition, perhaps more than anything else we have considered so far, it seems nearly impossible to give any remotely objective estimate of the expected value of pursuing these projects, so the case for prioritizing them will depend very much on 'squishy', subjective probabilities, with plenty of space for reasonable disagreement.[20] We find ourselves correspondingly very uncertain about the longtermist case for trying to permanently influence values or institutions.

millions of years on Earth, it will have achieved only a tiny fraction of its potential—an existential catastrophe. (Among other things, if humanity remains Earth-bound, it will survive for only a small fraction of its potential civilizational lifespan, i.e., will suffer 'premature' extinction.)

[19] Note that there are two distinct questions we can ask about proposed longtermist interventions: whether they beat neartermist benchmarks like the near-term benefits of bednet distributions, and whether they are (at least in some decision situations) *optimal* from a longtermist perspective, producing greater far-future expected value than any other available longtermist interventions. We focus primarily on the first question for two reasons. First, it seems plausible that any longtermist intervention that beats the neartermist benchmarks will be optimal by temporally impartial lights in at least some circumstances, where higher-priority longtermist interventions are not available or have already been carried out. (Among other considerations, the value of information favors testing out many promising strategies for improving the far future even if one, like x-risk mitigation, initially seems to have much greater expected benefits than the rest.) Second, the question of what longtermist interventions beat the neartermist benchmarks is especially relevant when our interest (as here) includes the robustness of the case for longtermism: if the list of such interventions is long and diverse, then even someone who is skeptical of interventions that are in fact (or by another's lights) more cost-effective might agree with longtermism on the basis of considering other interventions 'lower down the list'.

[20] As MacAskill (2022) notes, it is also true that a case for *not* prioritizing the projects in question would rest on equally 'squishy' subjective probabilities—probabilities that give projects of this nature higher far-future expected value than plausible near-term benchmarks don't seem actively outlandish. But there may be reason to err on the side of conservatism about the long-term impact of our actions when probabilities are squishy. In particular, perhaps we should start from a relatively confident prior that any given choice has only an extremely small effect on the probabilities of different long-term outcomes for humanity. Then, when our subjective estimates of the

2.6 Taking stock

The standard argument for longtermism aims to show most directly that *there exist some* interventions whose far-future expected benefits are many times larger than the highest available near-future expected benefits. Our focus in this section has been on *how large* is the class of interventions with that feature, beyond the handful of examples that appear in the existence proof.

In general, the thrust of our discussion is that (i) the majority of interventions for which there is a *fairly robust* such 'longtermist case' are in the category of mitigating technological x-risks, though (ii) that category is itself quite broad. We have found a couple of scattered interventions that are exceptions to this general rule (most plausibly, patient philanthropy and space settlement). Beyond technological x-risk mitigation and those scattered additional possibilities, the argument that any other interventions generate greater far-future benefits than the neartermist benchmark seems to us to depend much more heavily on subjective probabilities. This is of course not to say that no additional interventions with that 'longtermist' property exist, but it is to say that there is plenty of room for reasonable disagreement.

3 How many decision situations?

The decision situations that make longtermism most compelling are those in which some resource (e.g., money or work hours) can be allocated to work that directly mitigates technological x-risks. One is deciding, perhaps, whether to spend philanthropic funding on technological x-risk mitigation or instead on bednets, or whether to pursue a career in AI safety or instead in medicine.

But most real-world decision situations don't seem to belong to this category. When a transport minister makes decisions about an urban transport system, the possibility of taking the funding away from urban transport and donating it to AI safety research is not in the relevant option set: if they tried, the action would be blocked by higher authorities in the government or by the courts. When one is deciding what to have for breakfast, the available options include things like 'eat Cheerios' and 'eat peanut butter toast', not things like 'negotiate a nuclear arms reduction treaty': the latter aren't alternatives to the former.

It is *possible* that many of these other decision situations, too, are such that far-future considerations are the most important determinant of which options are best. But the usual argument for longtermism, focusing as it does entirely on cases of unconstrained resource allocation, does not establish that.

Beyond mere possibility: Are there any positive reasons to think that longtermism is true more broadly, encompassing many or even most real-world decision situations?

Here is one reason: Once it is agreed that philanthropic funding of technological x-risk mitigation has very high marginal benefits, any situation in which an agent's choice has financial implications for agents who are willing to fund such work inherits some of that

probabilities in a *particular* choice situation are based on ambiguous, non-robust evidence and arguments (and therefore more error-prone than our prior), we should update only very little from that prior. For a formal illustration of this idea, see Russell (n.d.).

longtermist significance. Suppose, for example, that someone on a monthly salary has committed to donate half the money left in her bank account at the end of each month to an organization working to prevent engineered pandemics. Then every purchasing decision she makes (where the options differ in price) has a fairly direct, though typically small, impact on total funding for x-risk mitigation. Similarly, every choice she makes at work and every choice in her personal life that affects her capabilities at work has an impact on her job performance, and thereby her expected future earnings, and thereby expected future funding for x-risk mitigation. Likewise, a policymaker facing any decision that affects the overall health of the economy thereby affects the financial wherewithal of many other agents, some of whom contribute to x-risk mitigation efforts.

The scope of this argument is somewhat unclear. For instance, most agents do not in fact donate any of their disposable income to x-risk mitigation, and so choices that affect their personal finances do not (at least in this way) implicate the total stream of longtermist funding. And perhaps most choices have no (or virtually no) financial implications for anyone—for instance, deciding where to sit at the family dinner table, or which park to visit with your children on the weekend. Even if most choices have *some* indirect impact on x-risk mitigation funding, this impact might be so attenuated that it is trumped by other factors, the standard argument for longtermism notwithstanding.

The proponent of expansive longtermism could respond here with what we might call the 'appeal to astronomical stakes'—the idea that the sheer scale of the long-term future is so astronomically great that *anything* that affects the probability of existential catastrophe by *just about any amount* is likely to outweigh any near-term considerations, even without crunching the numbers (and perhaps the same goes for other shifts in the probability of different long-term outcomes or trajectories, apart from existential catastrophe).[21] If this is so, then even highly indirect and heavily attenuated effects on this pool of funding may be overwhelmingly important.

On the other hand, one could easily accept the standard case for longtermism while believing that such claims about the marginal impact of x-risk mitigation funding are wildly overstated. If the marginal returns to x-risk mitigation are merely large and not astronomical (say, beating the relevant neartermist benchmarks by only 1–2 orders of magnitude, rather than 10–15, in the context of unconstrained philanthropic resource allocation), then the attenuated effects that our ordinary choices have on x-risk mitigation funding may well be trumped by neartermist considerations.

If there are more channels through which our everyday choices can influence the far future, however, those effects may be less attenuated and more significant. A more general version of the preceding argument from indirect effects on funding is (what we will call) the *many levers argument*.

> **Many levers argument:** There are many social, political, and economic variables whose present-day values make some not-completely-negligible difference to the expected value of the far future. For instance, there are reasons to think that a higher present rate of

[21] In this spirit, Bostrom (2013: 19) argues for the conclusion that 'the loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole' *without even considering any particular numbers* for the actual cost-effectiveness of available risk mitigation options.

> economic growth would be good in the long run and reasons to think it would be bad, and these reasons probably do not net out to exact neutrality. So in one direction or another, the present growth rate matters for the long-run future. The same might be true, for instance, of population size, average educational attainment, various cultural norms (e.g., norms of tolerance or civility), the relative power of particular states (that have a positive or negative influence on international affairs), and so on. Nearly every choice we make, however mundane, has *some* influence on *at least one* of these factors. For instance, if you eat a more expensive breakfast, you'll have less money to save, slightly reducing the expected economic growth rate; or if you eat a nutritious breakfast, you'll be more productive at work, slightly increasing the expected economic growth rate. And since the future is vast, even a small effect on a variable that has a small effect on the long-term trajectory of civilization can carry substantial expected value, positive or negative—enough to swamp the immediate stakes of, say, having a more or less delicious breakfast.[22]

As with the narrower argument concerning effects on existential risk mitigation funding, this argument relies to a significant extent on an appeal to astronomical stakes. Even if variables like GDP and population make some difference to the probabilities of different long-term futures, their effects may be indirect, and in many mundane decision situations, the strongest effect of the agent's choice on any of those variables may itself be indirect or simply very weak. But if the far-future stakes are astronomically large (say, equivalent to 10^{30} or 10^{52} lives; see fn. 27 below), then even these weak effects may be overwhelmingly important in expectation.

Our view is that this reasoning is suggestive but inconclusive. Even granting the standard argument for longtermism, it is still very much up for debate whether, say, the expected marginal benefit of philanthropic spending on x-risk mitigation is only a couple of orders of magnitude greater than neartermist benchmarks like anti-malarial bednets, or 10+ orders of magnitude greater, or something in between. And it is similarly unclear how much the ratio of long-term to near-term stakes is reduced in decision situations where our actions only affect important far-future outcomes very indirectly, but may affect important near-future outcomes directly. It is therefore extremely unclear how to assign the numbers required to flesh out the preceding arguments. If more rigorous assessment of these matters is possible, it would be very valuable.

4 How many resources?

Longtermism is often understood as having radical practical implications in the sense that it recommends a *very large* reallocation of resources from 'neartermist' interventions toward whatever would most effectively improve the far future in expectation. One starts to worry, perhaps, that there would be little left for dealing with the needs or preferences of the present; we today might all become slaves to the cause of optimizing the far future.

[22] This argument was suggested to us by Owen Cotton-Barratt (in conversation). Balfour (2021) makes an argument along similar lines for the claim that the long-term importance of x-risk has unpleasant totalizing implications, requiring that 'every waking moment' be governed by the imperative to minimize existential risk.

It might be helpful to make the issue more precise as follows. Suppose longtermism is true at the current margin. And imagine that we put more and more resources into the hands of agents who know that longtermism is (presently) true, and will spend those resources optimally.[23] Given a plausible model of the relevant diminishing marginal returns, at what point in this process, if ever, does longtermism cease to be true?

In this connection, three questions are pressing. First, how quickly would returns to investment in longtermist interventions diminish? Second, insofar as the amount of resources that would in an optimal scenario be *redirected* so as to better benefit the far future is large, how radical would the redirection be—in particular, how great a cost would it impose on the present generation? Third, if the optimal reallocation of resources is both large and radical, what should we make of this axiological fact in moral and practical terms?

First, then: How quickly would returns to investment in longtermist interventions diminish? For tractability, let us restrict the question to investment in existential safety (this will supply a lower bound on the continued cost-effectiveness of longtermist interventions more generally). There has been some investigation of the functional relationship between investment in a given area and marginal output in that area. For instance, Nicholas Rescher (1978; 1997) has proposed a 'law of logarithmic returns' to describe the growth of scientific knowledge. According to Rescher, knowledge increases logarithmically with the total quantity of resources invested in the scientific enterprise. In a series of blog posts, Owen Cotton-Barratt (2014a; 2014b; 2014c) has argued for more general versions of the same principle. Cotton-Barratt claims that (i) the *expected* returns to resources invested in solving a problem of unknown difficulty (e.g., figuring out how to prevent extinction-level pandemics or align powerful AI systems with human values) and (ii) the *actual* returns to resources invested in a problem *area* containing many independent problems of widely varying difficulty and importance are both approximately logarithmic within some central range (i.e., for resource investments that are not very small or very large). Cotton-Barratt's arguments could be reasonably taken to apply to spending on the mitigation of particular existential risks, or existential risk in general.

The case for this model is, as it stands, far from conclusive. But in the absence of any better-defended model, let us nonetheless consider its implications. If investment in particular cause areas obeys a law of logarithmic returns, it turns out that the optimal level of spending on x-risk mitigation or other longtermist cause areas depends very heavily on whether the current expected benefits of interventions in those areas are astronomical or merely large, in comparison to neartermist benchmarks. As a stylized illustration, suppose we can spend money on just two things: present consumption and x-risk mitigation. Assume that spending in either area yields logarithmic returns. Further assume that we currently spend 0.001% of world GDP (~$1 billion/year) on x-risk mitigation, and that the present marginal value of additional spending on x-risk mitigation is 10 times greater than the marginal value of transferring consumption to people in extreme poverty (who live on ~$500/yr). From these assumptions, we can conclude that the optimal level of spending on x-risk mitigation is less than 1% of world GDP.[24] On the other hand, if the present marginal

[23] Where are those resources coming from, i.e., at the expense of what other people or projects? We can imagine that, for instance, we simply give the longtermists newly printed bills or newly extracted precious metals, so that the resources in the hands of other people and projects shrink at approximately the same rate, retaining their relative proportions.

[24] Let v_c give the value of (optimally) allocating a given fraction of world GDP to present consumption, and v_m give the value of (optimally) allocating a given fraction of world GDP to x-risk mitigation. On the assumption

value of additional x-risk mitigation is 10,000 times greater than the marginal value of consumption for the world's poorest, the same reasoning tells us that it would be optimal to spend ~71% of world GDP on x-risk mitigation (leaving ~$3571.43 per person per year for present consumption).[25]

Our impression is that some longtermists believe that the marginal value of the best longtermist interventions exceeds the neartermist benchmark by 10:1 or less, while others believe that this ratio is 10,000:1 or more. And there is more than enough reasonable basis for this wide range of views. In particular, estimates of the expected future population that would be lost to existential catastrophe span tens of orders of magnitude.[26] Depending on how one distributes one's credence over these estimates, reasonable best guesses about the badness of existential catastrophe can easily span several orders of magnitude, at the very least. The upshot is that even the fairly strong assumption of logarithmic returns does not pin down an answer to the question of what portion of humanity's resources it is optimal to allocate toward longtermist causes.

Second: Assuming that a large reallocation of resources toward longtermist objectives is optimal, how radical would the optimal reallocation be compared to the status quo, and how great a cost would it impose on the present generation? To illustrate this question, consider the following two scenarios:

> ***Scenario 1*:** Most workers are employed in highly focused efforts to mitigate existential risks that have few if any co-benefits for the present generation. The largest employers are the asteroid detection and deflection industry, mathematical computing labs carrying out analysis directed toward AGI safety, and counterterrorism agencies working to contain threats from advanced biotechnology. A small proportion of the population works in food production, mainly supplying rice and beans to meet the basic caloric needs of the workforce. A small proportion works in education, but only to the extent needed to train up the next generation to continue the fight against existential risks. There is little funding for healthcare: generally it works out more cost-effective to let nature take its course, and to devote the resources that might fund healthcare instead to making the existential safety industry a little larger. Similarly for leisure, entertainment, hospitality, the arts, and games: these are neartermist luxuries that have no place in a society focused on minimizing existential risk.

that each individual has the same logarithmic function from consumption to utility, the optimal allocation of resources to present consumption will always distribute consumption equally. Allocating 1/25 (0.04) of world GDP to (equalized) consumption would give everyone in the world ~$500/yr. So, from the assumptions in the main text, we know that $v_c'(0.04) \approx v_m'(0.00001)/10$. If $v_c(x)$ and $v_m(x)$ increase like $\ln(x)$, then their first derivatives decrease like $1/x$. This means that $v_c'(0.04) \approx v_m'(0.0001)$. And *this* means that the optimal allocation of world GDP between present consumption and x-risk mitigation, which must equalize the marginal values of investment in each area, must exhibit the same ~400:1 ratio of consumption to x-risk mitigation spending. Thus, optimal spending on x-risk mitigation will be slightly less than 0.25% of world GDP.

[25] Now, we know that $v_c'(0.04) \approx v_m'(0.00001)/10000$, which implies that $v_c'(0.04) \approx v_m'(0.1)$. Thus, the optimal allocation of world GDP between present consumption and x-risk mitigation will exhibit a ~2.5:1 ratio of x-risk mitigation spending to present consumption.

[26] For instance, Millett and Synder-Beattie (2017) estimate the badness of extinction at $1.6 * 10^{16}$ life-years, based on the assumption that humanity would otherwise maintain a population of 10 billion individuals for the next 1.6 million years. On the other hand, Bostrom (2013) suggests that a future spacefaring civilization could support at least 10^{54} life-years worth of subjective experience. The issue is discussed in depth by Newberry (2021).

***Scenario 2*:** Returns to *highly focused* efforts to mitigate technological x-risks (e.g., research on AI control and alignment, or negotiating nuclear arms reduction treaties) diminish quickly, such that the optimal level of investment in these efforts is only a small fraction of world GDP. Beyond a relatively modest investment in those focused efforts, while it remains highly cost-effective to invest in technological x-risk mitigation, the most effective such investments also carry large near-term cobenefits. It is important for x-risk mitigation, for example, that there be effective and rational governance and decision-making at the national and supranational levels, a well-functioning economic system to implement society's preferences, and a populace that is free, prosperous, well-educated, and happy, which makes it a fertile generator of new ideas that will improve the long-run future, prone to support large cooperative projects, and immune to dangerous and destabilizing forms of radicalization.

From the partial point of view of the present generation, Scenario 1 is a bleak vision in which impartial concern for the far future has taken the joy out of life today. Scenario 2, on the other hand, is not so radically different from the status quo (or rather, differs largely in ways that are *beneficial* to present people). It involves some minor reordering of existing priorities and some change in the rationales for existing projects, but to a reasonable first approximation, Scenario 2 resembles the world that someone focused exclusively on the near term might aspire to create.

Between these scenarios, we find the bleak Scenario 1 somewhat less plausible as a longtermist optimum. It seems likely to us that a world like Scenario 1 could be stable only if human psychology were dramatically different. With actual human psychology, even if longtermism once commanded universal assent, the demands that longtermist morality could realistically make before leading to burnout, backlash, and conflict would likely be severely limited.

Third: The possibility that the optimal shift of resources toward optimization for the far future might be very radical (as in Scenario 1) naturally raises a 'demandingness' concern about longtermism: does an outlook that accepts this axiological claim also hold that we are morally obliged to pursue the identified optimum? If so, is that unacceptably demanding?

Here as elsewhere, though, it is important to keep the distinction between axiological and deontic claims in sharp focus. If the implications of an axiological longtermist thesis *together with maximizing consequentialism* strike one as overly demanding, then (absent some other reason for doubting the axiological longtermist claim) the natural response is to reject maximizing consequentialism, not to revise one's axiology or one's empirical beliefs.

At the same time, however, it is not plausible that axiological matters are *altogether irrelevant* to the deontic and practical issues. A morally conscientious person who rejects the conclusion that we ought to donate 99% of our income to the world's poorest usually does not conclude that the axiological analysis of global poverty is morally or practically irrelevant.[27] Similarly, if as an axiological matter something like Scenario 1 would indeed be optimal, it is much more plausible that deontically *some significant movement toward* Scenario 1 is warranted than that none whatsoever is warranted.

[27] It is much more plausible, in light of the facts about global poverty, that one ought to donate some significant but modest portion of one's income—perhaps, one's 'fair share' (see, e.g., Miller 2011)—than that because the *optimal* level of donation would be too demanding, there is no requirement to support the global poor at all.

5 Summary and conclusions

The standard argument for longtermism appeals to a narrow range of examples: interventions *to mitigate technological x-risks* that are available *at the current margin* in a *small (if important) class of decision situations*. Many who are initially moved by this argument, though, often subsequently endorse a more sweeping longtermist view: what we have called 'expansive longtermism'. There are three dimensions of this strengthening, corresponding to the three clauses just italicized.

First, the expansive longtermist holds that the class of interventions whose far-future benefits exceed any attainable near-future benefits is large and diverse, containing many things besides narrowly focused efforts to mitigate technological x-risks. On this view, the role of the appeal to technological x-risks in the standard argument is that of example to prove an existence claim.

Second, the expansive longtermist holds that the class of decision situations in which the value of the best options is determined primarily by far-future considerations is very large, encompassing not only situations of unconstrained resource allocation (as in some cases of private philanthropy and top-level public budget-setting), but also just about every aspect of public, professional, and private life.

Third, the expansive longtermist holds that longtermism is not only true *now, at the current margin*, but would remain true even after very significantly more resources were directed toward mitigating technological x-risks (or improving the course of the far future more generally).

In all cases, our main purpose in this essay has been to highlight that there is an important gap between what standard defenses of longtermism establish on the one hand, and this expansive picture on the other. But we have also considered the question of how plausible the more expansive longtermist view is, on each of the three dimensions of expansion. Regarding the question of 'how many interventions', we highlighted (i) patient philanthropy and (ii) accelerating space settlement as particularly promising candidates for 'longtermist interventions' that are not obviously matters of mitigating technological x-risks. Regarding the question of 'how many decision situations', we outlined potential arguments for the more expansive view from (i) indirect effects on the funding of technological x-risk mitigation and (ii) the more general idea of 'many levers' by which our present choices influence the far future. We found both arguments suggestive, but, as they stand, inconclusive. Regarding the question of how large and radical a reallocation of resources longtermist reasoning would deem optimal, we highlighted the issues of (i) how fast returns to spending on improving the course of the far future would diminish and (ii) how costly the recommended reallocation would be in terms of present-day welfare as particularly important open questions. In all these cases, though, our attempts at initial contributions are only that, and there is significant scope for careful research to improve the state of the debate.[28]

[28] For feedback on drafts of this chapter, we are grateful to Bradford Saad, David Thorstad, Philip Trammell, and participants in work-in-progress seminars at the Global Priorities Institute and Institute for Future Studies. For research assistance, we are grateful to Toby Newberry.

References

Aird, M. (2020), 'Crucial Questions about Optimal Timing of Work and Donations', *Effective Altruism Forum*, https://forum.effectivealtruism.org/posts/LD3mNJ367tSMna6WR/crucial-questions-about-optimal-timing-of-work-and-donations

Alexander, S. (2022), '"Long-termism" vs. "Existential Risk"', *Effective Altruism Forum*, https://forum.effectivealtruism.org/posts/KDjEogAqWNTdddF9g/long-termism-vs-existential-risk

Althaus, D. and Gloor, G. (2019), 'Reducing Risks of Astronomical Suffering: A Neglected Priority', *Foundational Research Institute*, https://foundational-research.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/

Anthis, J. and Paez, E. (2021), 'Moral Circle Expansion: A Promising Strategy to Impact the Far Future', in *Futures* 130 : 102756.

Arrhenius, G. and Bykvist, K. (1995), 'Future Generations and Interpersonal Compensations: Moral Aspects of Energy Use', in *Uppsala Prints and Preprints in Philosophy* 21.

Aschenbrenner, L. (2020), 'Existential Risk and Growth', Global Priorities Institute – Working Paper (Global Priorities Institute, Oxford University).

Balfour, D. (2021), 'Pascal's Mugger Strikes Again', in *Utilitas* 33/1: 118–124.

Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future*, PhD thesis, Rutgers University.

Beckstead, N. (2019), 'A Brief Argument for the Overwhelming Importance of Shaping the Far Future', in H. Greaves and T. Pummer (eds.), *Effective Altruism: Philosophical Issues* (Oxford University Press), 80–98.

Benatar, D. (2006), *Better Never to Have Been: The Harm of Coming into Existence* (Oxford University Press).

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', in *Utilitas* 15/3: 308–314.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4/1: 15–31.

Chan, S. (1997), 'In Search of Democratic Peace: Problems and Promise', in *Mershon International Studies Review* 41/1: 59–91.

Christiano, P. (2013), 'Giving Now vs. Later', *Rational Altruist*, https://rationalaltruist.com/2013/03/12/giving-now-vs-later/

Cotton-Barratt, O. (2014a), 'How to Treat Problems of Unknown Difficulty', *Future of Humanity Institute*, http://www.fhi.ox.ac.uk/how-to-treat-problems-of-unknown-difficulty/

Cotton-Barratt, O. (2014b), 'Theory behind Logarithmic Returns', *Future of Humanity Institute*, http://www.fhi.ox.ac.uk/theory-of-log-returns/

Cotton-Barratt, O. (2014c), 'The Law of Logarithmic Returns', *Future of Humanity Institute*, https://www.fhi.ox.ac.uk/law-of-logarithmic-returns/

Cotton-Barratt, O. (2020), '"Patient vs Urgent Longtermism" Has Little Direct Bearing on Giving Now vs Later', *Effective Altruism Forum*, https://forum.effectivealtruism.org/posts/Eh7c9NhGynF4EiX3u/patient-vs-urgent-longtermism-has-little-direct-bearing-on

Cotton-Barratt, O. and Ord, T. (2015), 'Existential Risk and Existential Hope: Definitions', technical report (Future of Humanity Institute).

Cowen, T. (2007), 'Caring About the Distant Future: Why It Matters and What It Means', in *The University of Chicago Law Review* 74/5: 5–40.

Cowen, T. (2018), *Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals* (Stripe Press).

Greaves, H. (2024), 'Concepts of Existential Catastrophe', in *The Monist* 107/2: 109–129.

Greaves, H. and MacAskill, W. (this volume), 'The Case for Strong Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).

Häggström, O. (2016), *Here Be Dragons: Science, Technology, and the Future of Humanity* (Oxford University Press).

Häggström, O. (this volume), 'The Hinge of History and the Choice between Patient and Urgent Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).

Hoeijmakers, S. (2021), 'Introducing the Patient Philanthropy Fund', *Founders Pledge*, https://founderspledge.com/stories/introducing-the-patient-philanthropy-fund

Kahneman, D. and Deaton A. (2010), 'High Income Improves Evaluation of Life but Not Emotional Well-Being', in *Proceedings of the National Academy of Sciences* 107/38: 16489–16493.

Landesman, C. (1995), 'When to Terminate a Charitable Trust?', in *Analysis* 55/1: 12–13.

Longview Philanthropy. (2021), 'About Longview Philanthropy', https://www.longview.org/about

MacAskill, W. (2022), 'Are We Living at the Hinge of History?', in J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan (eds.), Ethics and Existence: The Legacy of Derek Parfit (Oxford University Press).
MacAskill, W. (2022), *What We Owe the Future* (Basic Books).
Miller, D. (2011), 'Taking Up the slack? Responsibility and Justice in Situations of Partial Compliance', in C. Knight and Z. Stemplowska (eds.), *Responsibility and Distributive Justice* (Oxford University Press), 230–245.
Millett, P. and Snyder-Beattie, A. (2017), 'Existential Risk and Cost-Effective Biosecurity', in *Health Security* 15/4: 373–383.
Mello, P. (2017), 'Democratic Peace Theory', in P. Joseph (ed.), *The SAGE Encyclopedia of War. 4, Social Science Perspectives* (SAGE Knowledge), 472–475.
Moller, D. (2006), 'Should We Let People Starve—For Now?', in *Analysis* 66: 240–247.
Myers, D. (2000), 'The Funds, Friends, and Faith of Happy People', in *American Psychologist* 55/1: 56–67.
Nanda, N. (2022), 'Simplify EA Pitches to "Holy Shit, X-Risk"', *Effective Altruism Forum*, https://forum.effectivealtruism.org/posts/rFpfW2ndHSX7ERWLH/simplify-ea-pitches-to-holy-shit-x-risk
Newberry, T. (2021), 'How Many Lives Does the Future Hold?', GPI Technical Report No. T2-2021 (Global Priorities Institute, Oxford University).
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
Pettigrew, R. (2024), 'Should Longtermists Recommend Hastening Extinction Rather Than Delaying It?', in *The Monist* 107/2: 130–145.
Posner, R. (2004), *Catastrophe: Risk and Response* (Oxford University Press).
Rees, M. (2003), *Our Final Century: Will the Human Race Survive the Twenty-first Century?* (Basic Books).
Rescher, N. (1978), *Scientific Progress: A Philosophical Essay on the Economics of Research in Natural Science* (Basil Blackwell).
Rescher, N. (1997), 'The Law of Logarithmic Returns and Its Implications' in D. Ginev and R. Cohen (eds.), *Issues and Images in the Philosophy of Science* (Springer Dordrecht), 275–287.
Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking).
Russell, J. (n.d.), 'Planning for Pascal's Mugging', https://philarchive.org/rec/RUSPFP-3
Sagan, C. (1994), *Pale Blue Dot: A Vision of the Human Future in Space* (Random House).
Tarsney, C. (2023), 'The Epistemic Challenge to Longtermism', in *Synthese* 201/6: 1–37.
Tarsney, C. (forthcoming), 'Expected Value, to a Point: Moral Decision-Making under Background Uncertainty', in *Noûs*.
Thorstad, D. (2022), 'Existential Risk Pessimism and the Time of Perils', Global Priorities Institute – Working Paper (Global Priorities Institute, Oxford University).
Trammell, P. (2021a), 'Dynamic Public Good Provision under Time Preference Heterogeneity: Theory and Applications to Philanthropy', Global Priorities Institute – Working Paper (Global Priorities Institute, Oxford University).
Trammell, P. (2021b), 'New Products and Long-term Welfare', unpublished working paper, https://philiptrammell.com/static/New_Products_and_Long_term_Welfare.pdf
Trammell, P. (2021c), 'Patient Philanthropy in an Impatient World', unpublished working paper, https://docs.google.com/document/d/1NcfTgZsqT9k30ngeQbappYyn-UO4vltjkm64n4or5r4/edit#
Yglesias, M. (2022), 'What's Long-Term about "Longtermism"?', *Slow Boring,* https://www.slowboring.com/p/whats-long-term-about-longtermism#footnote-anchor-1

19

What Would a Longtermist Society Look Like?

Owen Cotton-Barratt and Rose Hadshar

1 Introduction

A *longtermist perspective* is a perspective which assesses actions almost entirely on the basis of their expected impacts on the far future.[1]

Longtermist perspectives currently influence resources at the margin. Only a few actors make decisions based on longtermist principles, so they must look for the best marginal use of resources from a longtermist perspective, in a world where relatively few resources are allocated on these principles. No actors' decisions hinge on what is the best *overall* allocation of resources from a longtermist perspective, and this question has received little attention.

If longtermist perspectives came to shape whole societies, people would need answers to this. Given the huge ratio between present and potential future generations, would almost all resources be allocated towards the future, leaving the present generation destitute? If not, why not?

There are several reasons to care about these questions. For one thing, it simply seems like a natural topic to explore: nothing about longtermism directly implies that the ideas should only be taken seriously by a small fraction of people, and if longtermism is an important moral truth the opposite may be true. More pragmatically, sometimes the implications of a position shed light on the position itself. Insofar as we care about whether longtermism is correct, it may be valuable to understand its implications. Moreover a better understanding of the implications of longtermism at a societal scale could deepen our understanding of what longtermism should imply about resource allocation at current margins. Finally, the set of actors taking longtermist ideas seriously may continue to expand, and it is possible that at some point in the future there will exist a society which is longtermist in some meaningful sense. Thinking through the implications of longtermism at a societal level in advance seems likely to improve outcomes in these futures.

This essay offers some very preliminary steps in such thinking. We outline some different senses of what a 'longtermist society' might mean, and sketch what might be implied in each case. Our four sketches are:

- A *partially longtermist society*—in which longtermism is taken seriously but is not dominant;

[1] This more or less corresponds to 'strong longtermist' in the sense of Greaves and MacAskill (2021). We don't mean to say that short-term impacts are excluded from the analysis altogether, merely that if one attaches importance to the whole of the future, then longterm impacts (say 100+ years) are the dominant term.

Owen Cotton-Barratt and Rose Hadshar, *What Would a Longtermist Society Look Like?* In: *Essays on Longtermism*.
Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press.
© Owen Cotton-Barratt and Rose Hadshar 2025. DOI: 10.1093/9780191979972.003.0019

- An *implausibly strict longtermist society*—in which every person cares about nothing else;
- A *strict longtermist state*—where the state is strictly longtermist but its people have other values;
- *Imperfections in longtermist states*—guesses about the pitfalls practical attempts might face.

These are not exhaustive. They are intended to stimulate further discussion on what longtermist societies might look like, and to provide stylised examples which make analysis a little more tractable.

2 Definitions

We have characterised *longtermist perspectives*, but we are concerned with actors making decisions. We assume that given an actor and a choice-situation, the actor will weigh different considerations against each other and make a decision. The various considerations might represent prudential and impartial perspectives, or different moral theories, or heuristics they have developed. We do not specify the process by which considerations are weighed against one another.

We will say an actor is *longtermist* if the assessment of one or more longtermist perspectives, or heuristics derived from longtermist worldviews, are admitted as valid and significant considerations in decision-making.

We will say an actor is *strictly longtermist* if the assessment of one or more longtermist perspectives, or heuristics derived from longtermist worldviews, are the *only* significant considerations in decision-making.

We will say an actor is *partially longtermist* if they are longtermist but not strictly longtermist.

3 A partially longtermist society

Perhaps the easiest kind of longtermist society to imagine is a partially longtermist society, where the state is partially longtermist and many (but not necessarily all) of the individual members of society are partially longtermist—in the sense that they admit longtermist perspectives as valid and significant considerations in decision-making, alongside other considerations. In this society, we may assume that it is acknowledged in public discourse that impact on the far future is an important feature of our actions today, but this is not treated as overwhelming.

If we look at analogues for partially longtermist societies in today's world, we can observe that societies have come to value animal rights, the environment, or gender equality. It is today common to find societies where perspectives on these issues are admitted as real and significant considerations for what is right to do. Pragmatically, such societies appear to put significant resources into protecting the thing valued, but not to the exclusion of everything else. Presumably the same would be true of partially longtermist societies (and is true, to the extent that some societies are already fairly characterised as partially longtermist):[2] they

[2] Most people would agree that future people matter to some degree, and many states take actions which benefit future people, although this may not be their motivation for action.

would invest in the longterm future, but a good fraction of resources would still be allocated to the good of current people for their own sake.

There are two main ways we might expect a partially longtermist society to diverge from present-day societies. First, a partially longtermist society would stop doing some things which occur in present-day societies, because of the negative externalities of those things for the longterm future. Second, a partially longtermist society would allocate more resources to *longtermist goods*: broadly understood, activities which aim to directly protect the potential value of the longterm future. (The difficulty of identifying which activities are in fact helpful for the longterm future complicates the definition of longtermist goods, but if we just count those resources which were intended to help, the picture remains fairly simple.)

Many activities which have negative externalities for the longterm future already have negative externalities for current generations and the nearterm future—for example, bad educational outcomes for children, or climate change. In other cases, negative externalities for current generations are less obvious. A category which is of particular salience in the case of the longterm future is research activity driving technological risk, and a partially longtermist society might be expected to coordinate around avoiding some avenues of research (at least unless and until they can be pursued in a way that is recognised as robustly safe).

Longtermist goods could include:

- implementation work to minimise natural extinction risks;
- research into and implementation of strategies to robustly avoid existential risks from technology;
- technological progress, to minimise the period of exposure to natural risks;[3]
- research into which actions have positive or negative externalities for the longterm future;
- research into and implementation of strategies to ensure that society stays well-governed as generations pass;
- research on the relative priority of these different objectives.

We can think of the total resources dedicated to the longterm future as the sum of the resources actively allocated to longtermist goods. (A more holistic analysis might include the economic cost of forgoing the activities which are refrained from on this account, but we do not see how best to measure this in principle.)

Clearly, present-day societies already allocate some resources to this basket of activities. A more longtermist society would allocate more resources to direct attempts to help the longterm future than present societies do, and these resources would of necessity be diverted from other potential goods. So there would be trade-offs to make between the good of future people and current people in a partially longtermist society, but these trade-offs would be similar in kind to the budgetary trade-offs we see today between, say, foreign aid and national social spending, rather than anything more extreme.

[3] Aschenbrenner (2020).

What might the absolute allocation of resources to directly helping the longterm future be in a partially longtermist society? This might vary considerably according to the different spread of values in different societies. One way of thinking about resource allocation would be to look for relevant reference points. For instance:

- 2% of European Union's GDP (EUGDP) was spent on environmental protection in 2021 (Eurostat 2022).
- 2.4% of global GDP was spent on defence in 2020 (Stockholm International Peace Research Institute 2022).
- 9.83% of global GDP was spent on healthcare in 2019 (World Health Organization Global Health Expenditure 2022).

These are all goods which are about protecting something which is very seriously valued by society. The exact amounts spent are probably sensitive to both the degree the things are valued, and the opportunities available to purchase the goods. If securing the longterm future—another such good—were valued a comparable amount, perhaps we might expect to likewise see single-digit percentages of GDP spent on it, unless opportunities to purchase longtermist goods are much more restricted than for goods like healthcare or defence. Or if it were valued less, but comparable to rich-country spending on foreign aid, it might be less than one percentage point but more than a tenth of a percentage point of GDP.

In practice, a partially longtermist society might have major research programmes in academia and elsewhere; government departments dedicated to different longtermist goods; greater legal restrictions or regulatory safeguarding on certain activities which increase existential risk; and some different cultural norms and values to present-day societies. But it would still, we presume, allocate substantial resources to the good of present people, as in present-day societies, and for intrinsic rather than instrumental reasons.

4 An implausibly strict longtermist society

What is the most extreme allocation of resources away from present people that would be coherent in a longtermist society? To answer this we will imagine an *implausibly strict longtermist society*, where each individual person (as well as the state as a whole) is strictly longtermist: that is, longtermist perspectives are the *only* significant considerations in their decision-making. It seems implausible that any society would ever take such a unanimous and extreme stance; the stance itself is also arguably implausible, but that is not the focus of our analysis.

Even in such a society, there would still be provision of nearterm social goods (which don't aim to directly protect the potential value of the longterm future), for instrumental purposes. Significant resources would be required to support whatever the optimal level of work on longtermist goods turned out to be: people whose basic needs for food, health, and shelter are unmet are unlikely to be able to perform the work required. Supporting work on longtermist goods would likely include:

- having and raising children;
- educating people so they are well-equipped to tackle challenging research work;

- mechanisms to match people with whatever type of work they are best suited for;
- producing technology and built environments that facilitate productive work;
- providing communities, entertainment, therapy, or whatever is necessary for people to be psychologically healthy and able to productively work over long periods;
- providing food that is nutritious and satisfying enough to enable proper focus on work;
- providing housing and domestic goods which help people to rest and stay productive;
- providing healthcare to help people stay or become maximally productive;
- general good stable governance of society;
- . . . and so on.

Many of what we normally regard as consumption goods are present here in the service of increasing people's productivity. This would be true for the productivity of everyone who was doing, or could do, any productive work (such as the work in the list above), not just of those working directly on longtermist goods.

So how much of the total economy would be devoted to longtermist goods? The answer to this will depend on other details of the society. For an extreme example, a society of subsistence farmers must spend a large proportion of their GDP on producing food; whereas the modern world spends around 10% of GDP on the food system in direct costs (Van Nieuwkoop 2019). The allocation to longtermist goods in an implausibly strict longtermist society is very sensitive to other needs, since (by definition) the society will devote everything it can to longtermist goods. So we should expect that an implausibly strict longtermist society at today's technology level would allocate a much higher share of its resources to longtermist goods than an implausibly strict longtermist society would have done a thousand years ago.

Taking the present world economy as a starting point, one approach to informing our views would be to look at the contribution of different industries to GDP today. For example, we could note that the world spends around 10% of GDP on food, 5% on education, 13% on construction, and 10% on healthcare; collectively these add up to 38% of world GDP. Some of this spending may be unnecessary, but it is hard to imagine it falling below, say, 15% of world GDP without a radical reform of the economy. This might suggest 85% as an upper bound for spending on longtermist goods, beyond which the value of the far future in fact begins to decrease with further spending on longtermist goods. On the other hand, much of the work on longtermist goods is research or other highly skilled work. This might require substantially larger investments in education; it may also limit the fraction of the population who can productively work on longtermist goods. Across the Organisation for Economic Co-operation and Development (OECD), around 2.6% of GDP is spent on research and development (R&D). A lower bound for spending on longtermist goods is probably several times this (since many people *could* work in R&D but do not); say perhaps 10%.

Overall, this suggests that an implausibly strict longtermist society would spend somewhere between 10% and 85% of its GDP on longtermist goods. This analysis is extremely crude, could be somewhat wrong at either end, and still leaves a very large range of plausible values. But we offer it as a start to the conversation. At minimum we think this refutes a kind of naive analysis that might have suggested that given the relative population sizes of current and potential future people, almost no (say <1%) resources would be dedicated

to the welfare of current people—a substantial investment is mandated for purely instrumental reasons (at least in the current world economy).

5 A strict longtermist state

Let us now consider a somewhat more realistic scenario where most individuals in society are partially longtermist (they admit other considerations into decision-making), while the state is strictly longtermist (and only admits longtermist perspectives into decision-making).

We presume that the implausibly strict longtermist society sets a lower bound for the allocation of resources to the welfare of current people; a strictly longtermist state might allocate substantially more resources to this. But would it? How can we consider how this would play out?

A starting point is to think of people's preferences as constraining factors on the productive work they can produce on existential security. Insofar as unhappy people are less productive than happy people, even a state which was efficiently and exclusively optimising for the good of future people would spend significant resources on meeting the nearterm preferences of its current citizens. An analogy here is companies whose ultimate aim is to maximise shareholder profits spending large amounts of money on goods like high-quality office space, food, and entertainment for their staff. We see this in industries (such as technology) where the work is complex and hard for an overseer to judge the quality of; in these cases it becomes more important for companies to inspire their employees to something akin to intrinsic motivation. Much of the work in existential security has a similar character; we guess therefore that at least partially meeting people's preferences would be instrumentally helpful for the strictly longtermist state in achieving its goals. We should therefore expect that a strictly longtermist state would allocate resources to goods like the arts, leisure, and care for the elderly, to the degree that their productive citizens place value on these things.

Note that some goods might relatively directly increase current people's productivity, while others might be much more indirect. More comfortable beds might directly improve people's sleep quality which in turn directly and locally improves the quality of their work. A music festival might locally have the opposite effect, but there are circumstances where a longtermist state might still allocate resources to such things: perhaps music festivals contribute to a richer culture which in turn stimulates creativity, or helps people to reach deeper states of fulfilment and purpose. (As these examples may suggest, there is no reason that the state need provide these productivity-enhancing things directly, rather than leaving individuals to obtain them via markets. If the state wishes to tip the scales of consumption choices to account for the externalities they cause, they can do that via taxes and subsidies.)

Another way that people's preferences may act as a constraint on the state's actions is through the need to maintain legitimate authority. If the state's actions deviate significantly enough from the preferences of its people, that could create an incentive for citizens to attempt to change the government (e.g. to one that wasn't strictly longtermist). This could provide a lower bound on the amount of resources spent on the welfare of current people—although it's hard to say exactly what this would be, and might depend on details of the political situation.

The presence of these various constraints on the state's actions means that a strictly longtermist state would have an interest in what its citizens thought of longtermism. It would be to the state's benefit if the people both cared about the longterm future, and regarded it as legitimate that the state act in such a way as to represent the interests of the longterm future. This could lead to the state wanting to invest in things which make people more sympathetic to longtermist perspectives. If the state were dogmatically longtermist, this might mean pro-longtermist propaganda. If the state were longtermist in a way that was contingent on the longtermism being correct, this might mean better education for its people, to give them the tools to think through for themselves how much (and how) to value the longterm future, and general increases in welfare, to remove immediate demands on people's attention and give them more space to consider (and hopefully come to agree with) perspectives such as longtermism.

6 Imperfections in longtermist states

If some day there are states which are well described as 'strictly longtermist', it seems likely that they will have flaws which deviate from an idealised picture (just as no capitalist or communist society has fully lived up to the ideals of their respective approach); and some of the deviation may be predictable. Insofar as longtermist states could actually arise, and to the extent that people can influence whether they do, it seems important to consider the ways in which they could go wrong. In principle these deviations could also apply to partially longtermist states, but for simplicity in the below we restrict our discussion to strictly longtermist states only.

In reality, longtermist states will be boundedly rational. They therefore will not optimally allocate resources according to their own lights, and might err in lots of directions. A particular concern from the perspective of the states' citizens might be states which approach longtermism in a relatively naive way, failing to fully identify the relationship between the good of current people and productive work for the good of future people. As much of this essay has implied, a very extreme departure from the good of current people seems unlikely: the connection between goods like adequate food and housing on the one hand, and productive work on the other, is obvious. But more subtle departures from the optimal allocation towards the good of current people might be harder to avoid. Thinking purely about suboptimality from a longtermist perspective (and so not taking into account other perspectives from which these scenarios might be strongly undesirable), we can consider several examples of how this might happen:

- A longtermist state might overwork its populace, resulting in *lower-quality* work than ideal.
- A longtermist state might create a large status differential between those working directly on longtermist goods, and those who 'merely' service that work. This might have various different negative impacts on the longterm, including:
  - leading those working on longtermist goods to worse judgements because it is not easy for them to consider possibilities that could cause them or people they care about to be shifted to lower-status groups;

  - leading the lower-status groups to civil unrest because they are dissatisfied with their position in society.
- A state that dogmatically adopted a longtermist perspective might indoctrinate its citizenry, in a way that prevented them from adopting better perspectives more conducive to the ultimate longterm good.
- A longtermist state that asked for big sacrifices from its populace might reduce the inclination of other nations to establish longtermist states in a way that could be bad for longterm prospects (even if its own populace remained highly productive).

In these scenarios, boundedly rational states incorrectly fail to allocate sufficient resources to the good of current people, even though it is instrumentally good for them to do so. But we might also be concerned about cases where something that we regard as an important component of welfare is *not* instrumentally helpful, and so is rationally discarded by longtermist states. For example, it's conceivable that a longtermist state might choose not to fund palliative care, or choose not to support lengthy retirement (provided that the productive members of the society didn't themselves place high value on these goods). If that were the case, we (as a society with pluralistic values) might well wish to reject the establishment of longtermist states which would act in this manner, if such a rejection lay within our power.

Another possibility is that a longtermist state would see a need to establish an invasive police state, in order to reduce existential risks from malicious actors with access to powerful technologies (see Bostrom 2019). Would we wish to reject that, if we otherwise supported a longtermist state? This might be a question which anyone grappling with establishing a longtermist state would want to address.

An implication of both the bounded rationality of states and the imperfect correlation between preference satisfaction and productivity is that it could be valuable to do further work uncovering the relationship between various kinds of social welfare and the longtermist good. Practically, the more obvious and well-evidenced these connections are, the less likely it is that a boundedly rational longtermist state would fail to supply those other goods. Philosophically, identifying objects of value which do not in fact contribute to the longtermist good (and in particular, identifying where preference satisfaction diverges from productivity), would help people to assess which kinds of entity they are or are not happy to establish or empower.

7 Conclusions

The space of possible longtermist societies is enormous. We have only considered a handful of dimensions (among a vast number) along which they could vary, and have only scratched the surface in our analysis of those dimensions.

We have seen that even the most strictly longtermist of societies would invest significant resources in the welfare of present-day people, for instrumental reasons. This sets a lower bound on the investment any realistic longtermist society might make. A somewhat more realistic strictly longtermist state would need to invest significantly more again in the welfare of present-day people, both from the perspective of preserving their productivity and

in order to maintain legitimacy. And in the most realistically accessible of the scenarios we discuss, the partially longtermist society, longtermist goods are treated with a similar respect to environmental goods.

However, we should also expect that longtermist states in practice will be imperfect in their resource allocation even by longtermist lights. This provides a reason to investigate now the longtermist foundations for societal goods which are normally understood as not relating to longtermism, in order to reduce the chance that these goods are incorrectly overlooked in longtermist states. This could also help people to assess the desirability of strictly longtermist societies in the first place, by better understanding which goods would be in or out of scope.

Working through the implications of longtermism at a societal level is a natural extension of studies in longtermism, and we hope further such work will shed light on the theory and practice of longtermism as a whole.

References

Aschenbrenner, L. (2024) 'Existential Risk and Growth', GPI Working Paper (Global Priorities Institute, Oxford University).

Bostrom, N. (2019), 'The Vulnerable World Hypothesis'. *Global Policy*, 10/4, 455–476.

Eurostat. (2022), 'National Expenditure on Environmental Protection by Institutional Sector and as Percentage of GDP, EU-27, 2006–2021', https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:National_expenditure_on_environmental_protection_by_institutional_sector_and_as_percentage_of_GDP,_EU-27,_2006%E2%80%932021_(EUR_billion_and_%25_of_GDP).png. Last accessed 6 January 2025.

Greaves, H. and MacAskill, W. (2021) 'The Case for Strong Longtermism', GPI Working Paper No . 5-2021, https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf. Last accessed 6 January 2025.

Stockholm International Peace Research Institute. (2022), *SIPRI Yearbook 2022: Armaments, Disarmament, and International Security* (Oxford University Press).

Van Nieuwkoop, M. (2019), 'Do the Costs of the Global Food System Outweigh its Monetary Value?', https://blogs.worldbank.org/voices/do-costs-global-food-system-outweigh-its-monetary-value. Last accessed 6 January 2025.

World Health Organization Global Health Expenditure database. (2022), 'Current Health Expenditure (% of GDP)', https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS. Last accessed 6 January 2025.

20

Is Extinction Risk Mitigation Uniquely Cost-Effective?

Not in Standard Population Models

Gustav Alexandrie and Maya Eden

1 Introduction

In the coming century, humanity may face global catastrophic risks stemming from climate change, nuclear war, pandemics, and emerging technologies such as artificial intelligence (Häggström 2016; Ord 2020). Many interventions for reducing these risks are likely to be cost-effective by the light of standard cost-benefit analysis (Posner 2004; Shulman and Thornley, this volume). However, it has often been argued that, under a zero rate of pure time preference[1], special priority should be given to the subset of these interventions that most effectively reduce the risk of human extinction. In a widely cited passage, Parfit (1984: 453) introduces this line of argument by comparing three possible outcomes:

(i) No catastrophe occurs.
(ii) A catastrophe kills 99% of the existing world population.
(iii) A catastrophe kills 100%.

Insofar as human life is valuable, (i) is clearly socially better than (ii), which in turn is better than (iii).[2] But which of these two differences is greater in terms of welfare loss? Counterintuitively, Parfit and many others have argued that, although the welfare difference between (i) and (ii) is greater if only the current generation is considered, the welfare difference between (ii) and (iii) is greater if all generations are considered equally. The motivation for this is that, while any global catastrophe would lead to an immense welfare loss for the current generation, human extinction would *additionally* lead to an even greater welfare loss by irreversibly preventing all subsequent generations from coming into existence.[3]

[1] Adopting a zero rate of pure time preference amounts to not discounting future welfare. Note that this is fully consistent with discounting future *consumption* based on the expected rate of economic growth and the diminishing marginal utility of consumption. Surveys of the arguments for and against adopting a zero rate of pure time preference are found in Dasgupta (2008), Greaves (2017), and Groom et al. (2022).

[2] For views to the contrary, see e.g., Benatar (2008) and Pettigrew (2024).

[3] Parfit writes that the 'Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second' (Parfit 1984: 453–454).

Gustav Alexandrie and Maya Eden, *Is Extinction Risk Mitigation Uniquely Cost-Effective?* In: *Essays on Longtermism*.
Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press.
© Gustav Alexandrie and Maya Eden 2025. DOI: 10.1093/9780191979972.003.0020

Therefore, in Parfit's view, '[w]hat matters *most* is how we respond to various risks to the survival of humanity' (Parfit 2017: 436, emphasis added). This line of thought has been invoked in cost-effectiveness analyses of interventions that reduce the risk of human extinction posed by asteroids (Matheny 2007), climate change (Ng 2016), and pandemics (Millett and Snyder-Beattie 2017). We refer to it as *the long-run argument for prioritizing extinction risk mitigation* (or simply, 'the long-run argument').

An important assumption underlying the long-run argument for prioritizing extinction risk mitigation over other types of risk mitigation is that the welfare effects of human extinction would be permanent, whereas the welfare effects of a non-extinction catastrophe would not. More precisely, the argument assumes that, if a non-extinction catastrophe were to occur, humanity would have a good chance of eventually recovering. However, the likelihood of such recovery depends on people's fertility decisions, which in turn depend on economic and social factors. Understanding these factors is necessary for determining whether extinction would indeed be uniquely consequential in the long run, or whether some non-extinction catastrophes would have comparably persistent effects on long-run population and welfare levels (Ord n.d.a; this volume).

In this chapter, we explore how shocks to the size of the current population might affect long-run population levels, and what this implies for philanthropic priority setting. We start by introducing a theoretical framework for quantifying the undiscounted cost-effectiveness of risk reduction efforts. A heuristic implied by this framework is that the undiscounted cost-effectiveness of reducing the risk of a negative population shock is proportional to the ratio of lives lost in the long run (in percentage terms) to lives lost in the short run (in percentage terms).

In the remainder of the chapter, we assess the implications of various population models for the relationship between decreases in current population levels and long-run population levels. First, we discuss shocks that reduce the current population level, but that leave all other factors of production unaltered. We show that, for such shocks, the assumption that population levels eventually recover after any non-extinction shock is implied by the Malthusian model of fertility (Malthus 1798). Importantly, however, this assumption is *not* implied by models that take fertility choices to be primarily determined by social norms. Nor is it implied by the Barro-Becker model (Becker and Barro 1988; Barro and Becker 1989), which is the workhorse model for studying the economic determinants of modern fertility dynamics. Indeed, in our calibration of the Barro-Becker model, non-extinction shocks to current population levels can result in permanent drops in long-run population levels that are disproportionately larger than the size of the initial shock.

We then proceed by analyzing events that reduce both population size and other factors of production proportionally by the same amount. Given constant returns to scale technology, such events leave economic determinants of fertility choices unaffected and therefore result in a permanent, proportional reduction in the size of the global population. Interventions that save lives *and* increase the capital stock in equal proportion therefore have permanent effects in standard economic fertility models. Our undiscounted cost-effectiveness framework suggests such interventions could be as cost-effective as extinction risk mitigation. Moreover, a back-of-the-envelope calculation suggests that these interventions may be even more cost-effective than extinction risk mitigation provided that the determinants of population levels remain sufficiently stable far enough into the future. While these cost-effectiveness estimates should be interpreted with considerable caution, they nonetheless suggest that interventions other than extinction risk mitigation can have significant social impact on the long-run social welfare.

2 Outlining the argument

Let us for simplicity restrict our attention to the subset of interventions whose social impact primarily stems from their effects on the number of people or life years that are brought into existence (as opposed to their effects on people's quality of life at any given time). We refer to this as the set of *population-affecting* interventions. Note that population-affecting interventions include both lifesaving interventions (e.g., antimalarial bednet distribution or extinction risk mitigation) and non-lifesaving interventions (e.g., changing fertility norms or sustainably increasing the supply of natural resources) that may affect long-run future population levels.

A stylized version of the long-run argument for prioritizing extinction risk mitigation over other population-affecting interventions can be stated as follows:

> (P1) The social value of a population-affecting intervention is approximately proportional to how much it increases expected long-run population levels.
>
> (P2) Additional philanthropic spending on extinction risk mitigation increases expected long-run population levels more than additional philanthropic spending on any other population-affecting intervention.
>
> (C) Therefore, additional philanthropic spending on extinction risk mitigation is more socially valuable than additional philanthropic spending on any other population-affecting intervention.

The first premise, (P1), can be supported by the following two claims:

> **Generalized totalism:** Social value increases linearly in the number of good lives[4] that are brought into existence (by the same amount regardless of when they are brought into existence).
>
> **Astronomical stakes:** In expectation, the vast majority of all current and future lives are going to be lives that are lived in the far future, and these lives are in expectation going to be good.

The reasoning is simple: if the social value increases linearly in the number of good lives, but the set of lives in the far future is vastly larger in expectation than the set of current lives, then the social value of any population-affecting intervention must be largely determined by its effects on long-run population levels. Although there are strong arguments in favor of *Generalized totalism* and *Astronomical stakes*, there are also important counterarguments.[5] However, for the remainder of this chapter, we grant that (P1) holds.

[4] 'Good lives' here simply refers to lives that contribute positively to social welfare.

[5] *Generalized totalism* is implied by additively-separable social welfare criteria such as total utilitarianism and total prioritarianism (see e.g., Blackorby, Bossert, and Donaldson 1995; Spears and Zuber 2021, for arguments in favor of these views). Moreover, as Tarsney and Thomas (2020) show, even non-additive axiologies, such as average utilitarianism, rank-discounted utilitarianism, and variable value views, converge in practice to the recommendations of additive axiologies if there is a large enough 'background population' that is unaffected by choices. Arguments for *Astronomical stakes* are discussed in Bostrom (2003; 2013), Beckstead (2013), Ord (2020), and MacAskill (2022), whereas arguments against are discussed in Thorstad (2023), which partly extends models developed by Adamczewski (n.d.) and Ord (2020).

The second premise, (P2), can be supported by the following two empirical hypotheses:

Recovery: Any drop in current population levels that does not result in human extinction would only have a transitory effect and would therefore not change long-run population levels.

Priority of saving lives: The population-affecting intervention that most cost-effectively increases long-run population levels is a lifesaving intervention.

The idea is again simple: if population levels always recover after non-extinction changes and if the most cost-effective way of increasing long-run population levels is a lifesaving intervention, then extinction risk mitigation must be the most cost-effective way of increasing long-run population levels. This establishes (P2). In the rest of this chapter, we analyze whether *Recovery* and *Priority of saving lives* are consistent with standard population models.

We consider three different population models. The first model is a social determinants model of fertility choices. According to this model, families target an ideal family size that is determined by social factors, primarily related to desirable family dynamics. We argue that, in this model, *Recovery* is unlikely to hold.

Another model that we consider is the Barro-Becker model (Becker and Barro 1988; Barro and Becker 1989), which emphasizes the role of economic factors in fertility choices. In this model, changes in population affect the macroeconomic conditions in ways that may ultimately affect fertility rates. As deaths or transitory changes in fertility rates may have permanent effects on population levels in this model, *Recovery* does not hold.

A third model that we consider is the Malthusian model (Malthus 1798). According to the Malthusian model, population levels are constrained by the availability of natural resources. Of the models that we consider, this is the only one that unequivocally supports *Recovery*, which makes it the most likely candidate for supporting the long-run argument for prioritizing extinction risk mitigation. However, we show that this model does not necessarily support *Priority of saving lives*: in the Malthusian model, interventions that permanently increase the supply of natural resources can permanently increase steady state population levels.

Table 20.1 Illustration of which models imply *Recovery* and/or *Priority of saving lives*.

Model	Does the model imply that *Recovery* holds?	Does the model imply that *Priority of saving lives* holds?
Social determinants model	No	
Barro-Becker model	No	
Malthusian model	Yes	No

3 Undiscounted cost-effectiveness

3.1 A framework for quantifying undiscounted cost-effectiveness

In this section, we introduce a framework for quantifying the cost-effectiveness of different interventions from the perspective of a longtermist decision-maker that gives equal ethical weight to each generation. The key assumption is that policymakers behave myopically, which is suboptimal from the longtermist's perspective. As a result, the undiscounted cost-effectiveness of an intervention is related to the ratio of its long-term benefits (in percentage terms) and its short-term benefits (in percentage terms).

The policymakers' problem. Suppose that the world's policymakers maximize the expected value of some random variable U. In the baseline scenario, the value of U is some (good) value, U_0. However, there are n other possible bad events that could occur. Event i occurs with probability p_i and results in value $U_i < U_0$ for the policymakers. The expected value of U is therefore given by

$$\left(1-\sum_{i=1}^{n} p_i\right)U_0 + \sum_{i=1}^{n} p_i U_i.$$

The probability of each event i is endogenous, as it depends on the resources that the policymakers devote to averting it. This implies the existence of a function, C_i, such that

$$p_i = C_i(m_i)$$

where m_i is the amount of resources devoted to averting event i. We assume that C_i is twice differentiable, strictly decreasing ($C_i' < 0$), and strictly convex ($C_i'' > 0$). This reflects the plausible assumption that the marginal reduction in the probability of i from an additional unit of resources devoted to averting i is diminishing, as the best opportunities for risk mitigation are successively exhausted.

The policymakers' optimization problem is thus given by

$$\max_{\{m_i\}_{i=1}^{n}} \left(1-\sum_{i=1}^{n} C_i(m_i)\right)U_0 + \sum_{i=1}^{n} C_i(m_i)U_i$$
$$s.t. \sum_{i=1}^{n} m_i = m.$$

where m is an exogenously given amount of resources that the policymakers devote to averting bad events.

Assuming an interior solution in which some of the policymakers' resources are devoted to the mitigation of each risk, the first-order conditions of this optimization problem imply the existence of some $\lambda > 0$ such that

$$C_i'(m_i^*)(U_i - U_0) = \lambda \quad \text{for all events } i \tag{1}$$

where m_i^* is the amount of spending to reduce the probability of event i that is optimal from the policymakers' perspective. Equation (1) states that, when policymakers allocate their risk mitigation spending optimally, the policymakers' marginal benefit of reducing the risk of event i, represented by the left-hand side of (1), is the same for all events i. The economic intuition behind this is that optimizing policymakers always prioritize spending on those events for which risk mitigation provides the highest marginal benefit, which drives down the marginal benefit of further spending until the marginal benefits of all risk mitigation efforts are equalized.

The longtermist's problem. Consider a longtermist who cares more about future generations than the policymakers do. Rather than maximizing the expected value of U, the longtermist wants to maximize the expected value of some W, which is given by

$$\left(1-\sum_{i=1}^{n} p_i\right)W_0 + \sum_{i=1}^{n} p_i W_i.$$

Assuming that the longtermist only has a small amount of resources, their marginal benefit of reducing the risk of event i is given by $C_i'(m_i^*)(W_i - W_0)$. Since the optimality condition (1) can be restated as $C_i'(m_i^*) = \dfrac{\lambda}{(U_i - U_0)}$, one can substitute for $C_i'(m_i^*)$ to get the following expression for the longtermist's marginal benefit of reducing event i:

$$C_i'(m_i^*)(W_i - W_0) = \lambda \frac{W_i - W_0}{U_i - U_0}.$$

The longtermist's marginal benefit from averting event i is therefore proportional to the ratio $(W_i - W_0)/(U_i - U_0)$.[6] This ratio, which we refer to as the *long-term value ratio*, increases proportionally in the degree to which mitigating the risk of an event i is cost-effective from the longtermist's perspective. Henceforth, we will use 'cost-effectiveness' to refer to 'cost-effectiveness from the longtermist's perspective'.

In what follows, we interpret U as the expected number of current lives relative to the baseline U_0. So, for example, $U_i = 0.75$ captures that event i reduces the current population by 25% relative to the baseline U_0. Similarly, we interpret W as the expected total number of current and future lives relative to the baseline W_0.[7] The statement $W_i = 0.75$ thus captures that event i reduces the sum of current and future population levels by 25% relative to the

[6] An important complication that we ignore in this chapter is that the policymakers may take (their expectations of) the longtermist's funding decisions into account when making their own funding decisions (see Trammell (2021) for an analysis of public good provision when funders have heterogeneous time preferences). These concerns may be less pressing if the longtermist's actions are instead conceived of as advocacy work to convince policymakers to reallocate their funds.

[7] This simplification seems reasonable under *Generalized totalism* (see section 2 and fn. 5 for more details).

baseline W_0. Note that this interpretation implies the normalization that $W_0 = U_0 = 1$ and $W_j = U_j = 0$ for any near-term extinction event j.[8]

Insofar as humanity is expected to last for a long time, the current population constitutes a relatively small fraction of the lives that the longtermist cares about. The long-term value ratio of spending to reduce the risk of an event can thus be heuristically interpreted as the ratio of the lives lost in the long run (in percentage terms) to the lives lost in the short run (in percentage terms) if the event were to occur.

3.2 The long-run argument for extinction risk mitigation

The framework presented in the previous subsection can be used to formalize a stylized version of the long-run argument for prioritizing extinction risk reduction. The first thing to note is that the normalization ensuring that $W_0 = U_0 = 1$ and $W_j = U_j = 0$ (for any extinction event j) implies that the long-term value ratio of reducing the risk of near-term extinction is normalized to 1, that is, $(W_j - W_0)/(U_j - U_0) = 1$.

As noted in section 2, an important assumption underlying the long-run argument for prioritizing extinction risk mitigation is that of *Recovery*. This is the assumption that, as long as humanity does not go extinct, long-run welfare and population levels would eventually recover after a shock. Under this assumption, for any non-extinction event i, the short-run welfare loss from i would be *proportionally* worse than the corresponding long-run welfare loss, that is:

$$\frac{U_i - U_0}{U_0} < \frac{W_i - W_0}{W_0} \tag{2}$$

for all non-extinction shocks i. Since $(U_i - U_0) < 0$ and $U_0 = W_0 = 1$, inequality (2) can be rearranged to say that $(W_i - W_0)/(U_i - U_0) < 1$ for all non-extinction shocks i. The long-term value ratio for efforts that reduce non-extinction risks is therefore strictly less than 1 under the recovery assumption.[9] Thus, given the recovery assumption, it is more cost-effective to reduce the risk of human extinction than to reduce other risks. The next section explores whether the recovery assumption holds in standard population models.

[8] Since U and W are unique up to affine transformations, one can add constants to each utility function to ensure that $U_j = W_j = 0$ for any extinction event j, and subsequently scale the utility function by some positive constant to achieve $U_0 = W_0 = 1$.

[9] How *much* lower than 1 is the long-term value ratio of reducing non-extinction risks given the recovery assumption? The answer depends on the rate of population recovery and the number of generations that come into existence after the recovery. As the fraction of generations that come into existence after the recovery tends to 1, the long-term value ratio of reducing non-extinction risks goes towards 0.

4 Shocks to population levels

4.1 Long-run effects in three population models

The number of people who will exist in the future depends on the fertility decisions of their predecessors.[10] These decisions are, in turn, the result of economic and social factors. The economic factors include individual wealth and factor prices. Wealth and income determine the amount of resources that people can devote to child-rearing, as well as the standard of living that they can afford each of their children. Wages capture a component of the costs of raising children, which often requires a reduction in work hours.

To understand how population shocks change fertility decisions, it is useful to understand how they affect the economic environment. If 50% of the population suddenly disappeared, there would be roughly 50% fewer workers. This would mean that each surviving worker could produce output using twice as much capital. For example, each farmer would have twice as much land; each factory worker would have twice as many machines, etc. As a result, standard economic theory predicts that there would be increases in the marginal product of labor and (therefore also) in wage rates. Average wealth would also increase, as the ownership of the economy's capital stock would be distributed among fewer people. These predictions appear to be broadly in line with the historical evidence indicating that the Black Death—the proportionally largest population shock in European history—led to a rise in living standards for ordinary people in late medieval Europe (Jedwab, Johnson, and Koyama 2022).[11]

How would these economic changes affect fertility decisions? One possibility is that they would have no effect whatsoever. There is some debate in the academic literature about the importance of economic factors for fertility decisions (see, e.g., De Silva and Tenreyro 2020). Some argue that fertility rates are primarily determined by cultural factors, such as social norms for the ideal family size. We call this the *social determinants model*. In this model, a population shock may have a persistent, proportional effect on long-run population levels, as the changes in economic conditions leave fertility choices unaltered. A 50% drop in population would result in population levels that are lower by 50% indefinitely (or at least until there is a change in the underlying social determinants of fertility). The social determinants model therefore suggests that the long-term value ratio of reducing the risk of catastrophes of any size is equal to 1, and so extinction risk mitigation is neither more nor less cost-effective than the mitigation of smaller catastrophes.

There are, however, channels through which economic factors could plausibly affect survivors' fertility decisions. The Malthusian model is perhaps the most well-known model of the economic determinants of fertility (Malthus 1798; see Becker 1988 for a more modern account of the model). This model emphasizes the income effect: as people's income and wealth increase, they can afford to have more children. It also emphasizes that production is constrained by the (fixed) quantity of natural resources. (Note that, in line with Becker (1988), we use the term 'Malthusian' in a broad sense to describe any model in which

[10] Another important determinant of population is aging, see Kuruc and Manley (this volume) .

[11] As Jedwab et al. (2022) point out, however, it should be noted that there is some disagreement among economic historians about 'the degree to which the post–Black Death era was a 'golden age' for workers' (150), and 'the extent to which these developments were driven by demographics' (150).

population is limited by a binding natural resource constraint. Importantly, this is compatible with any level of average consumption in the steady state, depending on people's fertility preferences.)[12]

The Malthusian model supports the recovery assumption that underlies the long-run argument for prioritizing extinction risk mitigation. To see this, consider a Malthusian economy with replacement fertility, and imagine a sudden negative shock to the size of the current population. Such a shock would leave more natural resources to go around, leading to higher wealth and income per person. Because of the income effect, this would in turn lead to above-replacement fertility that would remain until the population level recovers to its original size. Therefore, the Malthusian model implies that reducing the risk of extinction has a higher long-term value ratio than reducing the risk of smaller population shocks, whose effects are temporary. (That said, as we will argue in the next section, the Malthusian model also highlights the possibility of other interventions that may be as cost-effective as extinction risk mitigation.)

Another way in which changes in economic factors may affect fertility is through the substitution effect. Because labor shortages lead to higher wages, people have an incentive to work more. One way to have more time to work is to have fewer children. As a result, people may decide to have fewer children after a shock that reduces the size of the population. If this were to happen, the shock would be amplified: for example, a 50% drop in population may lead to long-run population levels that are even less than 50% of what they would have been otherwise. In the case of such shocks, the long-term value ratio would be greater than 1. Given our theoretical framework, reducing the probability of such shocks would therefore be more cost-effective than extinction risk mitigation.

To assess the relative strengths of the income and substitution effects, we present a calibration of the Barro-Becker model (Becker and Barro 1988; Barro and Becker 1989). The Barro-Becker model is the workhorse model for studying economic determinants of modern fertility dynamics. In this model, all capital is reproducible, so long-run population levels are not constrained by fixed natural resources. Since people are assumed to get utility both from consumption and from having children, the model allows for both the income effect and the substitution effect. The model and our calibration of it, which uses standard parameter values, are detailed in the appendix.

The results of our calibration of the Barro-Becker model are illustrated in Figure 20.1. The continuous line in Figure 20.1A plots the relationship between the size of the initial population shock and the resulting drop in steady state population levels, as implied by our calibration. For initial population shocks that are relatively small (<13%), the eventual drop in steady state population levels is proportionally smaller than the initial shock. However, for (non-extinction) initial population shocks that are relatively large (>13%), the reverse is true. In other words, our calibration implies that the relative strength of the substitution effect compared to the income effect increases in the size of the initial shock.

The continuous line in Figure 20.1B translates this relationship into undiscounted cost-effectiveness by plotting the long-term value ratio associated with reducing the risk of a

[12] In contrast, Malthus (1798: 40) himself seems to have held the narrower view that steady state consumption must be at the level of subsistence, as '[t]he passion between the sexes has appeared in every age to be so nearly the same that it may always be considered, in algebraic language, as a given quantity'.

shock against the size of that shock.[13] The long-term value ratio is below 1 when population levels recover from the initial shocks due to the income effect, and above 1 when the initial shock is amplified due to the substitution effect. Interestingly, our calibration implies risks that result in roughly a 35% drop in the size of the initial population are those with the highest long-term value ratio. We take these results to suggest that there could indeed be risk mitigation efforts for which the long-term value ratio is greater than 1. However, given the uncertainty associated with the model and its parametrization, we caution readers from drawing any conclusions stronger than this based on our calibration.

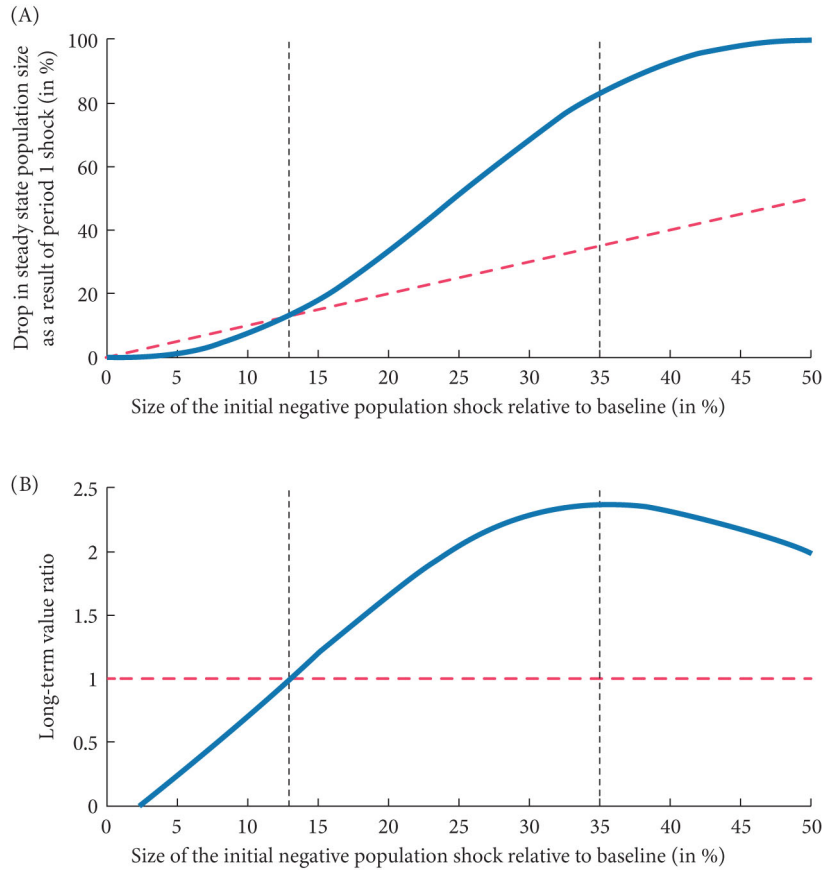


Figure 20.1 The continuous lines plot the drop in steady state population levels (A) and the long-term value ratio (B) against the size of the initial population shock, as implied by our calibration of the Barro-Becker model. The dashed lines represent the case where the percentage drop in the initial population level is the same as the percentage drop in steady state population levels.

[13] To make this translation we assume that $(U_i - U_0)$ is the proportional drop in initial population size, and that $(W_i - W_0)$ is the proportional drop in steady state population levels. Thus, Figure 20.1A plots $(W_i - W_0)$ against $(U_i - U_0)$, whereas Figure 20.1B plots $(W_i - W_0)/(U_i - U_0)$ against $(U_i - U_0)$.

4.2 How plausible is the Malthusian model?

So far, we have illustrated that the recovery assumption that underlies the long-run argument for prioritizing extinction risk mitigation is supported by the Malthusian model, but not by the social determinants model nor by the Barro-Becker model. In particular, the social determinants model suggests that extinction risk mitigation is neither more nor less cost-effective than the mitigation of smaller risks, and our calibration of the Barro-Becker model suggests that extinction risk mitigation might be less cost-effective than some other risk mitigation efforts. To evaluate the long-run argument, it is therefore of particular interest to further assess the plausibility of the Malthusian model, which does support the recovery assumption.

The first thing to note is that the Malthusian model is broadly considered irrelevant for explaining modern fertility dynamics. Capital accumulation and technological progress generated by industrialization have vastly increased the efficiency by which natural resources are utilized to the point that they no longer place binding constraints on population levels. Moreover, contrary to the predictions of the Malthusian model, fertility has fallen substantially in modern economies since the industrial revolution while wealth and income per capita have grown.

The case for nonetheless considering the Malthusian model is that Malthusian population dynamics may reemerge in the long run. First, evolutionary pressures for higher fertility might increase long-run population growth to the extent that natural resource constraints become binding once more (cf. Bostrom 2004; Collins and Page 2019).[14] Second, the development of artificial intelligence might result in rapid prolonged growth of machine labor and reproducible capital that eventually hits binding resource constraints such as energy or land (cf. Hanson 2016: 162–166; Korinek and Stiglitz, 2018: 383–386). In either of these cases, non-extinction shocks to population would only have temporary effects on long-run population levels, so the long-term value ratio associated with reducing such shocks would be lower than for extinction risk mitigation.[15] That said, it should be noted that fertility rates are currently below replacement in many high- and middle-income countries, suggesting that the long-run trend could be towards population decline rather than population increase (Basten, Lutz, and Scherbov 2013; Geruso and Spears, this volume). Moreover, in the next section, we argue that the Malthusian model suggests that there might exist interventions other than extinction risk mitigation with a long-term value ratio of 1.

5 Shocks to all factors of production

5.1 Theoretical considerations

The previous section focused on what population models imply about the relationship between long-run population levels and shocks that reduce population while leaving other

[14] In contrast, Arenberg et al. (2022: 2010) argue that 'intergenerational transmission of fertility is not sufficient to prevent longrun population decline'.

[15] It could also be argued that, even if one thinks that Malthusian future scenarios are unlikely, these scenarios are disproportionately important from a longtermist perspective (because some of these are the scenarios that contain the most social value) and could therefore nonetheless dominate expected social welfare calculations.

factors of production unaltered. In the Malthusian model, such shocks have no long-run effect at all, but in other models they may have proportional or even disproportional long-run effects. However, there is no reason to restrict attention to population shocks that leave other factors of production unaltered. Many shocks that affect population size also affect other factors of production. For example, wars result in human casualties, but also in the destruction of factories and cities. Similarly, climate change is likely to result in a large loss of lives, but also in a loss of natural resources. By considering the possibility of shocks that affect all factors of production, we can arrive at the more robust conclusion that, at least in our cost-effectiveness framework, there are interventions that are as cost-effective as extinction risk mitigation.

The key insight underlying this conclusion is that the economic factors affecting fertility are invariant to the scale of the economy. Population models typically assume that parents' decisions depend on their own wealth and income, but not on how many other people there are. Given constant returns to scale technology, per capita wealth and per capita income are not determined by the scale of the economy, but by the *ratio* of capital to labor. This feature implies that changing the scale of the economy, i.e., proportionally changing all factors of production, has no effect on fertility decisions. Consequently, these models imply that any shock that proportionately changes labor and capital would have a permanent effect on population levels.

Consider, for instance, a scenario in which a nuclear war kills 50% of the population *and* destroys 50% of the capital stock. Since the capital-labor ratio would be unaltered, constant returns to scale production technology implies that wages and rental rates would also be unaltered.[16] Economic fertility models would then typically imply that people choose to have the same number of kids as they would otherwise have had, which implies that the population size would remain permanently 50% lower. It follows that the long-term value ratio associated with reducing the risk of such shocks is 1. Therefore, our theoretical framework suggests that the undiscounted cost-effectiveness of reducing these risks is as high as that of reducing the risk of human extinction.

One might question whether this hypothetical possibility is empirically relevant. For example, perhaps there are no available interventions for reducing the likelihood of shocks that would proportionally reduce all factors of production. Alternatively, perhaps, for whatever reason, policymakers behave less myopically when it comes to shocks that proportionately affect all factors of production. To address these concerns, we propose a concrete intervention and provide a back-of-the-envelope calculation indicating that, if the Barro-Becker model (in which all capital is reproducible) holds indefinitely, our proposed intervention is more cost-effective than extinction risk mitigation.

5.2 Back-of-the-envelope calculation

Consider an intervention that saves lives and proportionally increases the stock of reproducible capital.[17] Further, assume that reproducible capital can substitute for natural resources

[16] An important caveat that we ignore here is that population levels may well have important effects for the rate of technological progress, as emphasized by, e.g., Jones (2022).

[17] Note that standard economic growth models with exogenous population imply that increases in population lead to increases in the marginal product of capital, which in turn means that more investments are made,

in production, so that the Barro-Becker assumption of constant returns to scale is plausible in the long run.

Distributing bed nets in malaria-prone regions in low-income countries is considered a highly cost-effective way of saving lives. According to a recent estimate by GiveWell, it costs around US$5,000 to save a life by distributing bed nets (GiveWell 2022).

To maintain a constant capital-labor ratio, this intervention must be accompanied by a proportional change in the capital stock. In other words, the value of the global capital stock must be increased by the current value of the capital stock per person. Importantly, the returns to this capital must accrue to the people whose lives were saved, as their future fertility decisions depend not only on their wages but also on their wealth. An implementation of our intervention would be a combination of distributing bed nets in a malaria-prone region while simultaneously transferring wealth to that region, either as direct transfers or through investment in infrastructure. Note that the wealth transfer would have to consist of resources that would otherwise have been consumed rather than invested.

To calculate the required capital investment, note that, according to the World Bank data for year 2021, global gross domestic product (GDP) per capita is around $12,000. About two-thirds of this is attributable to labor income, suggesting that capital income per capita is around $4,000. The standard no-arbitrage condition for investment implies that the marginal product of capital is equal to the interest rate plus the depreciation rate, $r + \delta$. A reasonable parameterization is $r = \delta = 0.05$, so that $r + \delta = 0.1$. Because capital income is equal to the marginal product of capital multiplied by the capital stock, i.e., $\$4{,}000 = 0.1 \cdot k$, the per-capita capital stock, k, is given by:

$$k = \$40{,}000.$$

This suggests that the combined intervention of saving a life and increasing the capital stock to offset the decline in the capital-labor ratio would cost around $45,000. It is notable that the bulk of the cost is the capital investment component rather than the lifesaving component.

This estimate suggests that, with $100, it is possible to save $\$100/\$45{,}000 \approx 0.0022$ of a life while maintaining the capital-labor ratio constant. Given a current world population of around 8 billion, this constitutes a permanent, proportional increase in the population of about $0.0022/(8 \cdot 10^9)$. Using Greaves and MacAskill's (2021) estimate that the expected number of future lives is around 10^{24}, it follows that the total number of lives saved in the long run by spending $100 on our proposed intervention is

$$\frac{0.0022 \cdot 10^{24}}{8 \cdot 10^9} = 2.75 \cdot 10^{11}.$$

While estimates such as the one above should not be interpreted literally (see Karnofsky 2011), it is nonetheless worth noting that the estimated returns of our proposed intervention are substantially higher than the returns that Greaves and MacAskill (2021) estimate for extinction risk mitigation. According to their estimates, the expected number of lives

eventually pushing the capital-labor ratio back to the steady state. These dynamics work differently in the Barro-Becker model, which is why we instead envision an intervention that both saves lives and increases the capital stock.

saved from spending $100 on asteroid detection is 300,000 and the expected number of lives saved from spending $100 on biosecurity is $2 \cdot 10^8$. Hence, astonishingly, compared to their biosecurity estimate, the back-of-the-envelope calculation above suggests that our proposed intervention saves around 1,000 times more lives for the same amount of money.

There are, of course, many extremely simplifying assumptions that go into the back-of-the-envelope calculation above. Importantly, it assumes that the Barro-Becker model holds indefinitely, which seems questionable given that future technology may allow for very different modes of reproduction. Technological developments could for instance potentially have substantial effects on fertility decisions and population growth by facilitating sex selection (Kolk and Jebari 2022), cloning (Saint-Paul 2003), and perhaps even mind-uploading (Hanson 2016). Moreover, combining lifesaving with capital investment admittedly amounts to an unconventional and perhaps politically impractical intervention.

Our aim here is not to argue that our estimate is reliable or that our proposed intervention is among the most effective ways of improving long-run welfare. Our aim is rather to illustrate, using an empirically grounded example, that there may indeed be interventions other than extinction risk mitigation that are cost-effective in virtue of having a long-run effect on the size of the global population.

5.3 Increasing the stock of natural resources

Although the back-of-the-envelope calculation above assumes that reproducible capital can substitute for natural resources in production, a conceptually similar calculation could in principle be performed in the case of the Malthusian model where reproducible capital and natural resources are not substitutes. In the Malthusian model, permanently increasing the supply of natural resources by some proportion would (via the income effect) increase long-run population levels by the same proportion. For example, if the quantity of arable land constrains long-run population levels, then our cost-effectiveness framework implies that preventing permanent destruction of arable land would increase short-run and long-run social value by the same proportion, and therefore have a long-term value ratio of 1—the same as extinction risk mitigation.

Similarly, if we anticipate humanity (or its descendants) to eventually become an intergalactic civilization, there is a resource constraint in the form of the number of reachable galaxies. For each year that intergalactic space expansion is delayed, $(2 \cdot 10^{-8})\%$ of the reachable universe is permanently lost due to the exponentially accelerating expansion of the universe (Ord n.d.b: 23; also cf. Armstrong and Sandberg 2013). The Malthusian model therefore suggests that speeding up intergalactic space expansion would have a permanent effect on population levels in the far future. If there are interventions that would accelerate intergalactic space expansion without having any short-run benefits, the long-term value ratio of these interventions could be greater than that of extinction risk mitigation.[18] However, it is worth noting that previous literature addressing this question generally finds

[18] Note that this conclusion only holds under the rather strong assumption that the myopic policymakers spend *some* positive amount of resources on these space expansion interventions (perhaps because they care about the far future a little bit). Without this assumption, the policymakers' problem has a corner solution, which violates the conditions we used to derive the long-term value ratio in subsection 3.1.

that extinction risk mitigation is much more cost-effective than speeding up space expansion (Bostrom 2003; Ord n.d.b).

6 Conclusions

Assessing the long-run argument for prioritizing extinction risk mitigation is important for philanthropic priority setting. Drawing on standard economic fertility models, the present chapter poses a challenge for this argument. As illustrated in section 5, such models typically imply that any shocks that proportionally decrease all factors of production have proportional, permanent effects on long-run population levels. Therefore, in our theoretical cost-effectiveness framework, the undiscounted cost-effectiveness of mitigating such shocks is comparable to that of extinction risk mitigation. Moreover, a back-of-the-envelope calculation, using the Barro-Becker model and plausible empirical estimates, implies that our proposed intervention, which combines bed net distribution with wealth transfers, is more cost-effective than extinction risk mitigation (provided that the Barro-Becker model holds indefinitely). Although these cost-effectiveness estimates are mainly intended to serve as helpful illustrations and should therefore be interpreted with considerable caution, they nonetheless suggest that interventions other than extinction risk mitigation could have significant impact on long-run social welfare.

In addition, our analysis of pure population shocks in section 4 highlights the possibility that the most cost-effective interventions might be those that mitigate large, non-extinction catastrophes rather than those targeted at extinction risk mitigation. The reason for this is that some reasonable fertility models have nontrivial long-run dynamics: a large, non-extinction shock to population may be amplified in the long-run. However, more work is needed to assess the likelihood of such amplification as well as possible ways to mitigate shocks of this kind.

Our challenge to the long-run argument for prioritizing extinction risk mitigation is thus not that extinction risk mitigation is less cost-effective than the argument purports, but rather that there may exist other interventions that are equally or perhaps even more cost-effective. In particular, we point out that (i) the argument seems to rely on the assumption that humanity would eventually recover after any shock to population, and (ii) the recovery assumption is violated by standard population models with the important exception of the Malthusian model.

Our challenge is particularly pressing if fertility rates are expected to remain low in the long run or if the exogenous rate of human extinction is expected to be high. In both of these scenarios, long-run population levels are unlikely to be governed by Malthusian dynamics. Conversely, the long-run argument for prioritizing extinction risk mitigation seems more resilient to our challenge if evolutionary or technological factors are expected to result in large future population levels limited by binding natural resource constraints. However, as argued in subsection 5.3, there might be interventions that have higher undiscounted cost-effectiveness than extinction risk mitigation even in these Malthusian scenarios.

Our discussion also provides insights about the potential long-term effects of different *types* of global catastrophes. In particular, it suggests that the extent to which a catastrophe destroys capital is an important factor for assessing recovery dynamics. Asteroids or wars, which result both in deaths and in the destruction of capital, are likely to have very different

long-run population effects compared to pandemics, which could result in the same numbers of deaths while leaving the capital stock largely intact. Our results point to the possibility that the former type of catastrophe may lead to a proportional loss in long-run population levels, whereas the latter type of catastrophe may result in either a disproportionately large or a disproportionately small long-run population effect. This suggests that assessing the potential of catastrophic events to destroy reproducible capital and natural resources—in addition to their potential to cause fatalities directly—is of special significance for long-term social welfare.

Acknowledgments

We want to thank Loren Fryxell, Luis Mota Freitas, Michael Geruso, Rossa O'Keefe-O'Donovan, Caroline Falkman Olsson, Kevin Kuruc, Charlotte Siegmann, Dean Spears, Luca Stroppa, Benjamin Tereick, and Phil Trammell for helpful discussion and feedback on previous drafts of this chapter.

Appendix

This appendix introduces the Barro-Becker model in more mathematical detail (see Becker and Barro 1988; Barro and Becker 1989, for further discussion) and describes how we calibrated the model to generate Figure 20.1.

Time is discrete and infinite, and each generation is alive in one time period only. The utility of each person alive in period t is

$$u_t = c_t^{\sigma} + \alpha\, n_t^{1-\epsilon} u_{t+1}$$

where $c_t > 0$ is the consumption of a person in generation t; $\sigma \in (0,1)$ captures the marginal utility of consumption, $n_t \geq 0$ is the number of children; $\alpha > 0$ is a parameter that governs how much people value having children; $(1-\varepsilon) \in (0,1)$ captures the marginal utility of having children; and u_{t+1} is the expected utility of each child.

People allocate their labor incomes, w_t, and their capital incomes, $(1+r_t)k_t$, between their own consumption, child rearing expenses, and saving for the benefit of their children. The budget constraints are thus given by

$$w_t + (1+r_t)k_t = c_t + n_t(a(1+g)^t + bw_t) + n_t k_{t+1}.$$

The cost of raising children is given by $a(1+g)^t + bw_t$. The first component of the cost, $a(1+g)^t$, is a cost in terms of goods, which is assumed to grow at the rate of technological progress. The second component of the cost, bw_t, is a time cost: each child requires sacrificing a fraction $b \in (0,1)$ of the individual's work time.

Output is produced using constant returns to scale technology in capital and labor. Technological progress is constant and labor augmenting. Output at time t is given by

$$Y_t = A(N_t k_t)^{\zeta} \cdot ((1+g)^t N_t (1-bn_t))^{1-\zeta}$$

where g is the rate of technological progress, A is the baseline productivity level, and N_t is the population size in period t, and ζ is the capital intensity parameter. In this model, the interest rate, r_t, and the wage rate, w_t, are determined based on the marginal products of capital and labor, respectively.

In our calibration, we focus on a steady state without technological progress $g = 0$, reflecting the hypothesis that, in the long run, the stock of knowledge will converge, as new ideas will get increasingly harder to find, and substantial resources will have to be devoted towards maintaining the stock

Table 20.2 Calibration parameters

Parameter	*Description*	*Value*
δ	Capital depreciation rate	0.72
σ	Intergenerational substitutability parameter	0.3
ε	Diminishing returns to number of children	0.288
a	Material cost of childrearing	11.4
b	Time cost of childrearing	0.16
α	Intergenerational altruism parameter	0.09
ζ	Capital intensity of production	1/3

of knowledge. The length of a period is taken to be 25 years, roughly corresponding to the age of fertility. The depreciation rate, δ, is chosen based on an annual depreciation rate of 5% a year, roughly in line with global averages. The capital intensity parameter, ζ, is chosen to roughly match the long-run capital income share.

The preference parameters, σ and ε, and child-rearing cost parameters, a and b, are taken directly from the calibration in Cordoba (2015). The parameter α is calibrated to generate a steady state with constant population. The productivity parameter, A, is normalized to match average wages, which are specified in 2010 dollars (this is not an important normalization). Given these parameter values, the Barro-Becker model implies the relationship between initial shock size and steady state population drop implied by Figure 20.1.

References

Adamczewski, T. (n.d.), *The Expected Value of the Long-Term Future* (unpublished manuscript).

Arenberg, S. et al. (2022), 'Heritable Fertility Is Not Sufficient to Prevent Long-Run Population Decline', in *Demography* 59/6: 2003–2012.

Armstrong, S. and Sandberg, A. (2013), 'Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox', in *Acta Astronautica* 89:1–13.

Barro, R. J. and Becker, G. S. (1989), 'Fertility Choice in a Model of Economic Growth', in *Econometrica* 57/2: 481–501.

Basten, S., Lutz, W., and Scherbov, S. (2013), 'Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility', in *Demographic Research* 28: 1145–1166.

Becker, G. S. (1988), 'Family Economics and Macro Behavior', in *The American Economic Review* 78/1: 1–13.

Becker, G. S. and Barro, R. J. (1988), 'A Reformulation of the Economic Theory of Fertility', in *The Quarterly Journal of Economics* 103/1: 1–25.

Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future*, PhD thesis, Rutgers University.

Benatar, D. (2008), *Better Never to Have Been: The Harm of Coming into Existence* (Oxford University Press).

Blackorby, C., Bossert, W., and Donaldson, D. (1995), 'Intertemporal Population Ethics: Critical-Level Utilitarian Principles', in *Econometrica* 63/6: 1303–1320.

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', in *Utilitas* 15/3: 308–314.

Bostrom, N. (2004), 'The Future of Human Evolution', in Bostrom, N., Ettinger, R. C. W., and Tandy C. (eds.), *Death and Anti-death: Two Hundred Years after Kant, Fifty Years After Turing* (Ria University Press), 339–371.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4/1: 15–31.
Collins, J. and Page, L. (2019), 'The Heritability of Fertility Makes World Population Stabilization Unlikely in the Foreseeable Future', in *Evolution and Human Behavior* 40/1: 105–111.
Cordoba, J. C. (2015), 'Children, Dynastic Altruism and the Wealth of Nations', in *Review of Economic Dynamics* 18/4: 774–791.
Dasgupta, P. (2008), 'Discounting Climate Change', in *Journal of Risk and Uncertainty* 37/2: 141–169.
De Silva, T. and Tenreyro, S. (2020), 'The Fall in Global Fertility: A Quantitative Model', in *American Economic Journal: Macroeconomics* 12/3: 77–109.
Geruso, M. and Spears, D. (this volume), 'Depopulation and Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
GiveWell. (2022), *Our Top Charities*. https://www.givewell.org/charities/top-charities (accessed 31 December 2022).
Greaves, H. (2017), 'Discounting for Public Policy: A Survey', in *Economics & Philosophy* 33/3: 391–439.
Greaves, H. and MacAskill, W. (2021), 'The Case for Strong Longtermism', GPI Working Paper No. 5-2021 (Global Priorities Institute, Oxford University).
Groom, B., Drupp, M. A., Freeman, M. C., and Nesje, F. (2022), 'The Future, Now: A Review of Social Discounting', in *Annual Review of Resource Economics* 14: 467–491.
Häggström, O. (2016), *Here Be Dragons: Science, Technology and the Future of Humanity* (Oxford University Press).
Hanson, R. (2016), *The Age of Em: Work, Love, and Life When Robots Rule the Earth* (Oxford University Press).
Jedwab, R., Johnson, N. D., and Koyama, M. (2022), 'The Economic Impact of the Black Death', in *Journal of Economic Literature* 60/1: 132–178.
Jones, C. I. (2022), 'The End of Economic Growth? Unintended Consequences of a Declining Population', in *American Economic Review* 112/11: 3489–3527.
Karnofsky, H. (2011), 'Why We Can't Take Expected Value Estimates Literally (Even When They're Unbiased)', *The GiveWell Blog*, https://blog.givewell.org/2011/08/18/why-we-cant-take-expected-value-estimates-literally-even-when-theyre-unbiased/(accessed 31 December 2022).
Kolk, M. and Jebari, K. (2022), 'Sex Selection for Daughters: Demographic Consequences of Female-Biased Sex Ratios', in *Population Research and Policy Review* 41: 1–21.
Korinek, A. and Stiglitz, J. E. (2018), 'Artificial Intelligence and its Implications for Income Distribution and Unemployment', in Agrawal A., Gans J., and Goldfarb A. (eds.), *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 349–390.
Kuruc, K. and Manley, D. (this volume), 'The Ethics, Economics, and Demographics of Delaying Aging', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
MacAskill, W. (2022), *What We Owe the Future* (Basic Books).
Malthus, T. (1798), *An Essay on the Principle of Population As It Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Goodwin, M. Condorcet and Other Writers* (J. Johnson, in St. Paul's Church-Yard).
Matheny, J. G. (2007), 'Reducing the Risk of Human Extinction', in *Risk Analysis: An International Journal* 27/5: 1335–1344.
Millett, P. and Snyder-Beattie, A. (2017), 'Existential Risk and Cost-Effective Biosecurity', in *Health security* 15/4: 373–383.
Ng, Y. K. (2016), 'The Importance of Global Extinction in Climate Change Policy', in *Global Policy* 7/3: 315–322.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books).
Ord, T. (this volume), 'Shaping Humanity's Longterm Trajectory', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Ord, T. (n.d.a), *The Value of Very Long-reaching Effects*. Unpublished manuscript.
Ord, T. (n.d.b), 'The Edges of Our Universe', arXiv:2104.01191, https://arxiv.org/pdf/2104.01191.pdf (accessed 31 December 2022).
Parfit, D. (1984), *Reasons and Persons* (Oxford University Press).
Parfit, D. (2017), *On What Matters*: Volume 3 (Oxford University Press).
Posner, R. A. (2004), *Catastrophe: Risk and Response* (Oxford University Press).
Pettigrew, R. (2024). 'Should Longtermists Recommend Hastening Extinction Rather Than Delaying It?', in *The Monist* 107/2: 130–145.
Saint-Paul, G. (2003), 'Economic Aspects of Human Cloning and Reprogenetics', in *Economic Policy*, 18/36: 73–122.

Schulman, C. and Thornley, E. (this volume), 'How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).

Spears, D. and Zuber, S. (2021), 'Foundations of Utilitarianism Under Risk and Variable Population', IZA Discussion Paper No. 14515, https://ssrn.com/abstract=3879363 (accessed 31 December 2022).

Tarsney, C. and Thomas, T. (2020), 'Non-additive Axiologies in Large Worlds', GPI Working Paper No. 9-2020 (Global Priorities Institute, Oxford University).

Thorstad, D. (2023). High risk, low reward: A challenge to the astronomical value of existential risk mitigation. *Philosophy & Public Affairs*, *51*(4), 373-412.

Trammell, P. (2021), 'Dynamic Public Good Provision under Time Preference Heterogeneity: Theory and Applications to Philanthropy', GPI Working Paper No. 9-2021 (Global Priorities Institute, Oxford University).

21
Depopulation and Longtermism

Michael Geruso and Dean Spears

'Future people count. There could be a lot of them. We can make their lives go better.' This is Will MacAskill's elegant introduction to longtermism for a popular audience in *What We Owe the Future* (MacAskill 2022: 9). It is the starting point of an argument for prioritizing the wellbeing of the seemingly endless stream of future people. Or, more specifically, people who may exist if humanity can evade the nearer-term existential risks that threaten it. In this chapter, we consider an important other possibility: There might not be a lot of them, after all.

MacAskill's book contains a striking diagram: If you have the book, look at page 15. A series of small stick figures, each representing 10 billion people, stretches for five pages. Nine-hundred and fifty-four stick figures dramatize the possible future of human lives yet to be lived. (MacAskill reports that, if he were not abbreviating to save paper, he would ideally include 5 million figures over 20,000 such pages.) He computes this number as the straightforward multiplication of 10 billion people living on Earth for each of 500 million years.[1]

We hope that so many excellent lives happen. If they do, it would mean that the global human population had stabilized at 10 billion (or, perhaps, at some other size, over a different length of future time). We make the case that stabilization is a future worth working for in our book *After the Spike* (2025). This chapter references and builds on facts and arguments in *After the Spike* to bring evidence from population science into dialogue with longtermists. The most important fact is this: The population science community projects that depopulation will happen, not global stabilization.

What could close off the possibility of so many future lives? Here is one way: that people continue, for a few centuries, to have the sort of birth rates that are now normal for most of humanity. With a high likelihood, the human population size will begin shrinking soon, within the lifetime of children alive today. What happens after that is less certain, but sustained depopulation is a likely possibility. Two-thirds of people now live in a country where fertility rates are not high enough, on average, to prevent depopulation.[2] By depopulation, we mean that the global population declines generation after generation. The decline would be approximately exponential if fertility rates are stable below two. The dividing line that

[1] MacAskill is not the only philosopher endorsing longtermism, and his are not the only arguments and explanations, but it is hard to overstate the traction that this presentation has received as a representation of one of Longtermism's core ideas. At the time of our writing, the Wikipedia (2023) entry for Longtermism had exactly two diagrams—both variations on the same idea as MacAskill's stick figure diagram, these produced by Our World in Data (Roser 2022). Newberry (2021) presents similar computations in a working paper. Later in *What We Owe the Future*, in the contexts of economic stagnation and value lock-in, MacAskill discusses the possibility of depopulation. Here we offer an expanded treatment of this possibility—its trends, its causes, and its consequences—drawing on our research in economics and demography.

[2] See details in Spears and Geruso (2025).

Michael Geruso and Dean Spears, *Depopulation and Longtermism* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Michael Geruso and Dean Spears 2025.
DOI: 10.1093/9780191979972.003.0021

separates population growth from depopulation is whether fertility is consistently above or below two children per adult woman, on average globally. Below that critical threshold, the next generation will not replace the last. Nearly all rich countries are already below two. The most populous poor and middle-income countries are as well. And in the places where fertility rates are still above two today, fertility is well below its past peak and falling.

To give a concrete numerical example: Consider the US, where fertility is not especially low by rich-country standards: 1.6 children per woman. Europe, China, Japan, and a set of other countries together amounting to 38% of the world population all already have fertility below this level. What if the whole world converges to the fertility rate that is normal in the US today? (At the risk of over-emphasizing, this would mean an *increase* in fertility in many countries where fertility is already lower than in the US.) How many stick figures would be needed to represent the count of all future human lives in that case? How many of MacAskill's 5 million figures would remain?

Three figures. If the whole world reaches and sustains a fertility rate like the US has now, then there would be fewer than 30 billion future human births, ever (Spears et al. 2024). There have been about 120 billion human births so far, since the beginning of our species. So that would mean that humanity is now four-fifths over, only one-fifth remaining. This outcome would not require low fertility to be sustained for millennia, or even for more than a few centuries. By 2350 CE, there would be only 20 million births per year compared to around 140 million in 2022. The last time so few people were born was sometime in the 9th century. The Mayan civilization was waning then. The Vikings were just getting started.

Longtermism (understood broadly as a body of scholarly arguments focused on making things go well for future people, because there could be a lot of them) typically focuses on the risk that the future ends quickly, due to a disaster like a pandemic, a supervolcano, an asteroid, or losing a war with artificial intelligence (AI) (Ord 2020). This chapter encourages longtermists to include on their research agenda the neglected possibility of slower, sustained depopulation—generations of exponential decay in population size over a few centuries until there are only a few million or few hundred million of us. Whether fertility rates, after that time, are such that the population stabilizes or continues falling,[3] it might not require a huge disaster to close off our flourishing future.

The goal of this chapter is to bring population science and population economics into dialogue with the community of longtermists who are thinking about wellbeing into the far future. Our maintained assumption is that most longtermists would agree that, to eventually achieve a flourishing far future, it is valuable that over the coming few centuries a complex global economy endures and the number of people does not become small enough to be highly vulnerable to extinction from a threat that a larger population could sustain. We review population projections and other social scientific facts that show that fertility rates that are normal in much of the world today would cause population decline that is faster and to lower levels than is commonly understood, threatening the long-term future.[4]

[3] In Spears et al. (2024) we quantify, with coauthors Sangita Vyas and Gage Weston, the consequences for long-run population dynamics of various possible future transitions to replacement fertility. There we show that the ultimate steady state size of the global population depends substantially on when the transition to replacement fertility begins.

[4] Unfortunately, this chapter cannot do its job, within its word limit, and also be an essay about population ethics, gender inequality, climate policy, or expanding our moral circle to include potential people. But these issues are important to us. We have written about them in *After the Spike*, and we touch on them briefly here.

We proceed in four sections. Section 1 presents demographic projections. These projections take the UN's projection to 2100 and then extend it using standard tools from population science. The result is that, on the timetable of a few centuries, depopulation could happen very fast, even if global fertility rates are not far below two children per woman, on average.

Section 2 responds to the question 'How can you be so sure?' We are not. That uncertainty is itself important. Like all threats to long-term flourishing, depopulation is a risk, not a certainty. Raftery and Ševčíková (2023), for example, compute that there is a 90% chance that global fertility remains below the stable replacement level out to 2300. In other words, they forecast a 90% chance of sustained depopulation—which implies a 10% chance of a fertility reversal that would also reverse depopulation. The historical cases of modern, below-replacement-fertility populations offer precisely zero examples of rebound to sustained fertility rates high enough to prevent depopulation.[5] But we won't rule out a reversal. The uncertain possibility is sufficient to make understanding depopulation an urgent cause. Section 2 also addresses common objections like: *Won't fertility decline someday reverse, or eventually equilibrate to two children per woman or more?* And: *Won't unprecedented technological progress disrupt everything before we get to 2300, so none of your population science matters?* Maybe! But 'maybe' also suggests 'maybe not'. And there is evidence to draw on that can inform these conjectures.

Section 3 briefly tours the consequences of depopulation for longtermist goals. Could a much smaller global population sustain a complex, modern, information-based economy? We review arguments from macroeconomics and other social sciences that it might not; we take seriously that risk. None of our arguments require fertility rates to stay low forever to threaten longtermists' goals. Timing is often neglected in discussions of long-run population, as if any population path might lead us to the same bright future. But it shouldn't be, as we discuss in section 3. Whatever future longtermists hope for, they should not be confident that progress towards these outcomes would not be closed off by a much smaller population in the next several hundred years.

Section 4 asks about possible policy responses. Plausible responses would require a clear understanding of why fertility is declining basically everywhere, but nobody really knows yet. There are no shovel-ready solutions to reverse this phenomenon because the basic science is incomplete. We describe how, contrary to popular myth, fertility policy has rarely (and possibly never) been effective at making and sustaining large changes to population-level fertility rates.[6] In this sense, responding to depopulation is not yet tractable at the level of policy, even though the present neglect of the underlying scientific issues makes it particularly ripe for progress. Understanding of depopulation is today where climate science was a half-century ago: It was important then that scientists were measuring carbon

[5] Population sizes have, of course, risen and fallen through history due to large swings in *mortality*—wars, plagues, etc.—but fertility rates have almost never been low before the last century and have never rebounded from low to high.

[6] Perhaps this is surprising if you have not encountered it before. But to tell you what the history and social science says: Governments sometimes try to coerce people to have babies; governments sometimes try to coerce people not to have babies. It is typical, with such policies, to wreck people's lives, wreck the economy's human capital, and wreck society's compact between the governed and the government. And yet, the evidence is far from clear, despite the popular myths, that such coercive policies have managed to change fertility much from the course it would have followed without coercion. As we explain in *After the Spike*, nobody yet knows of a policy response that could actually do much to change fertility.

concentrations, recognizing the system-level challenges, and raising questions, even though they had neither the computing power to produce an integrated climate assessment model nor the technological foundations for a clean energy infrastructure. This chapter is intended as a rousing call to more research and better understanding of depopulation.

1 Living in strange times: history and projections

Why do we project the future to contain only a few more stick figures, while MacAskill would draw 5 million, each representing 10 billion lives? Because we are answering a different question. MacAskill's stick figures describe what could happen, if humanity overcomes all barriers to sustaining a population of 10 billion for 500 million years, until the Earth becomes uninhabitable because of changes in the Sun. His illustration is not intended to ask who would conceive, gestate, and parent so many babies.

Our demographic projections, in contrast, do ask a stylized version of this. Our computation uses a cohort-component-model projection from population science. Such a model quantitatively tracks each hypothetical birth cohort as they age, have children, and die off along the way.

With this model, we answer the question of what would happen if the world follows the UN demographers' medium projection until 2100, and then each country converges over the following few decades to a level of low fertility that is already common—for example, fertility like the average fertility of the United States in 2021 (or of South America, or Europe, or East Asia). Although we present our own long-term projections that we have made with coauthors (Spears et al. 2024), we are not the first to document what would happen if fertility rates stay low. Our results are broadly consistent with those of Basten, Lutz, and Scherbov (2013) and Raftery and Ševčíková (2023), who also quantify uncertainty out to 2300.

Figure 21.1 is the answer to our question of why the future could be so small.[7] We call it the Spike.[8] It plots the number of births in each year for a long time into the past and a short while into the future. Look, for now, at the solid line, which makes our focal, illustrative assumption of a future where global fertility is like US fertility in 2021: a total fertility rate (TFR) of 1.66—meaning 1.66 children per woman, on average. As we will verify in Table 21.1 below, from a zoomed-out, longtermist vantage, the example of 1.66 will be informative of anything in the ballpark. Whether we consider 1.66 or 1.8 (present South America) or 1.2 (present East Asia) simply does not matter for our conclusions.[9] It would be

[7] Figure 21.1 repurposes the projections from Spears et al. (2024), where they were first published after peer review; we are grateful for the support of our coauthors Gage Weston and Sangita Vyas in producing these projections. For earlier research consistent with our results, see Basten et al. (2013). Data for years before 1950 is our construction from Table 1 of Kaneda and Haub (2022).

[8] In our book *After the Spike*, and in the 18 September 2023 *New York Times* opinion article 'The World's Population May Peak in Your Lifetime. What Happens Next?', 'The Spike' is a plot of the global population, rather than the global annual count of births.

[9] 1.66, in particular, is not a prediction we are making: There is no reason to believe that future global fertility will just so happen to be numerically like the country where we happen to be writing in the most recent data year available when we happen to be writing. We simply take 1.66 children from the US in 2021 as a non-outlandish illustrative example: Out of nine women, for instance, one has no children, two have one, five have two, and one has three kids.

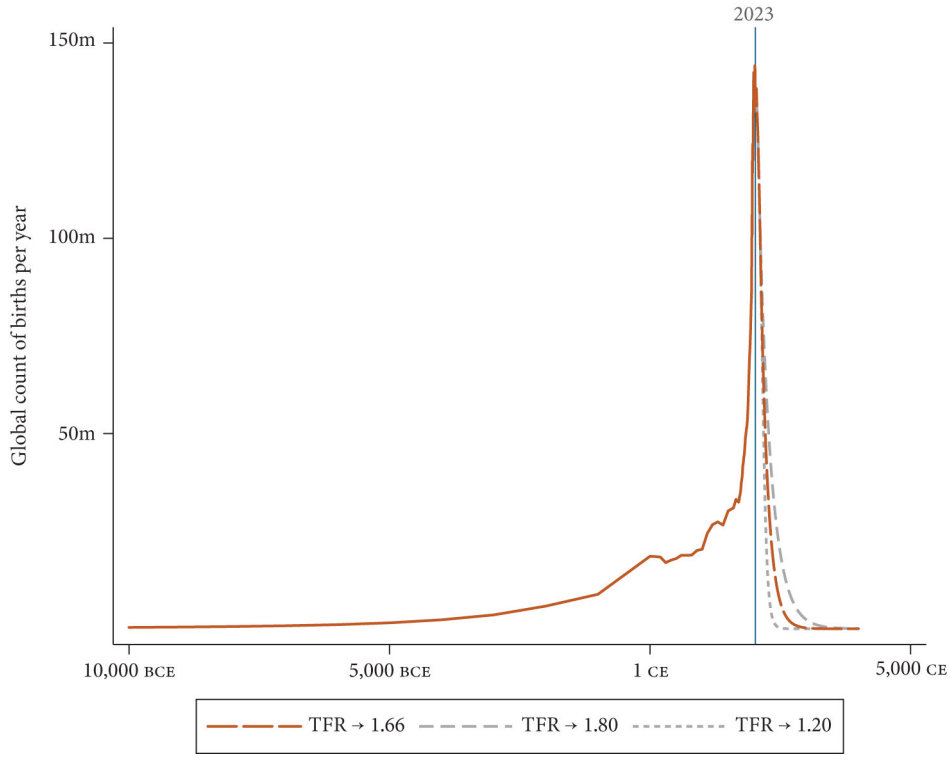


Figure 21.1 The Spike: If fertility rates stay low, then peak births per year has already passed.

Notes: Figure 21.1 plots historical estimates up to the present and cohort-component model projections for the future. Any such projections are conditional on a scenario for future fertility rates. The bold line assumes that future fertility converges to a TFR of 1.66; gray lines assume alternative below-replacement scenarios for future fertility, detailed in Table 21.1.

an error to focus too much on the particulars of any of these numbers and miss that they all imply the same broad pattern.

The area under the Spike describes the total number of human lives ever lived. Given the emerging new normal of low fertility, Figure 21.1 conditionally projects 150 billion lives will ever be lived. Of these, 120 billion have been born in the past.

If so, humanity is four-fifths over.

Of course, there is something silly about following the projection down forever. We do not think there is any chance that the math would hold until the last couple has only one child. Using this projection is merely a way of quantifying: *Humanity's numbers could quickly get small enough to be vulnerable to something bad.* And we show elsewhere (Spears et al. 2024) that what happens over the next couple of centuries could affect population sizes for a very long time, even if fertility rates someday recover to replacement level. Even if global fertility rates begin an unprecedented increase to 2 at some point in the 2100s, each one-year delay in this fertility rebound would cause a 0.7% decline in the permanent size of the stabilized, future population.

The Spike shows us that our times could be very strange, relative to the rest of human history, past and future. Although it may be hard to see, given the millenia-wide scale of Figure 21.1, we (you, the reader, and we, the authors) have already lived through the peak.

Total global births per year crested in 2012.[10] The world is already on the downward slope of the Spike.

The population size peak will happen after the births peak, but that is in our near-term future as well: The expert opinion of the population science profession is that, in a few decades, global fertility is very likely to fall below an average of two births per woman. That triggers a shrinking population. The UN projects a global total fertility rate of 1.86 in 2100, as its central estimate, and believes that the size of the world population will peak in the 2080s. Wolfgang Lutz and colleagues at the International Institute for Applied Systems Analysis (IIASA; Lutz, Butz, and Samir 2014) project 1.67 babies per woman in 2100 and place the peak in the 2070s. The Institute for Health Metrics and Evaluation (IHME) at the University of Washington (2020) projects a similar 1.66 for 2100 and a peak in the 2060s. The authors of these various studies and reports have good reason to care about the fine details that separate their work from one another's. But for our purposes here, these numbers are no different from each other: Each group expects a future of fewer than two[11] babies per two adults within the coming decades.[12]

Table 21.1 provides another illustration of what would happen if fertility falls below two, as demographers predict, and if it stays below two, under various assumptions.[13] The columns offer five scenarios of hypothetical asymptotic fertility (corresponding to the dashed lines in Figure 21.1). The bottom row of Table 21.1 reports the year in which the world would again have as few as 10 million births per year under each scenario. Ten million is an arbitrary marker,[14] meant to help us imagine a scenario in which economic and technological complexity might, we imagine, face meaningful constraints or risks—low enough, we propose, to be a risk to longtermist goals. The last time there were only 10 million births per year was around 800 BCE (Kaneda and Haub 2022), about the same time as what may be the first surviving account of a sundial (in the book of Isaiah)—before the first instructions to make glass appear in cuneiform tablets, before cast iron, and before stirrups.

An important lesson of Figure 21.1 and Table 21.1 is that change could come quickly. The population would shrink rapidly, to less than 10 million births per year, in just a few centuries

[10] According to the UN's historical records and central projection, for as long as they project, there will never again be as many babies born in a year as there were in 2012. (United Nations 2022)

[11] Why is two children per woman—or a little more than that; 2.05 to 2.1 in low mortality settings—the important dividing line between exponential growth and exponential decay? Because, when artificial sex selection is absent, human biology yields about 100 female births for every 105 male births. If, on average, 100 adult females produce fewer than 205 births (plus one or two more for the few babies in a low-mortality population who will not survive to reproductive ages) then they have not reproduced themselves (100 females) in the next generation. So, when fertility is above two per woman, each generation is larger than the last. When fertility is below two per woman, each generation is smaller than the last. And the rules of exponential growth and decay govern the population size, so growth or shrinkage compounds across generations. (In higher mortality settings, where a non-negligible fraction of females die before middle-age, replacement fertility can be higher—for example, in excess of three.)

[12] This is the advantage of counting in the coarse units of 10-billion-birth stick figures. If you disagree with our particular open-ended decay, then your alternative equations to end the model and ours would have to differ by 5 billion births not to agree on the rounded count of figures.

[13] In 2011, Gietel-Basten, Sobotka, and Zeman (2014) solicited the opinions of expert population scientists about a plausible long-run asymptotic fertility rate. This exercise settled on 1.75, which is within the range of our Table 21.1. We conjecture that this subjective expectation would be even lower today, after 12 further years of subsequent fertility decline. Raftery and Ševčíková (2023) project statistical distribution with a median global total fertility rate of 1.72 for both 2250 and 2300, strikingly in line with Gietel-Basten's estimate.

[14] Ten million is an arbitrary round number which turns out to be about 7% of the number of births that will occur this year. If you think the number that matters instead is twice or half as many as our focal 10 million example, ok. All the logic here still applies, just a handful of decades sooner or later.

Table 21.1 The robustness of our conclusions to alternative future fertility rates.

Hypothetical asymptotic fertility	**1.8**	**1.66**	**1.5**	**1.2**	**1.0**
Example 2023 country or region, according to UN	**South America**	**US**	**Europe**	**East Asia**	**South Korea**
% of 2023 world population in countries at or below this fertility rate	43%	38%	25%	19%	1%
% of all human lives which would have been already born	77%	82%	85%	86%	86%
% of all human lives which would remain yet to be born	23%	18%	15%	14%	14%
Number of future stick figures (future births ÷ 10 billion, rounded)	3	3	2	2	2
Year birth count falls below 20 million	2495	2345	2275	2240	2235
Year birth count falls below 10 million	2660	2445	2345	2280	2270

under any of these possible global total fertility rates. So fertility does not have to remain below two forever to be a threat to longtermist goals—merely for these next few centuries.[15]

What is striking about both Figure 21.1 and Table 21.1 is that it makes little difference to the shape of the Spike or to the final number of stick figures in humanity's future whether one assumes that fertility will converge to what is now normal in the Americas or to what is now normal in Europe or East Asia. For some challenges, like strained social welfare systems due to inverted age pyramids, the difference between a TFR of 1.4 (Japan) and 1.7 (US) is massive. But for the question of how many humans may yet be born, anything much below 2 leads to a very similar end: Only 20 or 30 billion lives yet to be lived.

2 How can we be so sure?

You may be asking: How can we be so sure that global fertility rates will fall below 2 and stay there for centuries?

Our first answer is: We are not sure, of course! No one should pretend to be sure that destructive AI will be invented or that an asteroid will cross Earth's path or that the supervolcano below Wyoming will erupt, nor when. But longtermists take these risks seriously. We should take the uncertainties of low fertility seriously, too.

Our second answer is that, even though we cannot be sure, the evidence of social science is aligned with two demographic facts. One fact is that falling fertility is found in essentially every population and subpopulation, even places with different economic, social, and policy environments. 'Falling' here does not mean *towards* two children per woman (the dividing line between exponential growth and decay); falling means right through two and below. The other fact is that, empirically, fertility rates that have fallen and stayed well below replacement levels have so far never rebounded to and stabilized at levels that would avoid depopulation.

Declining fertility is nothing new: Even as the population size has been growing (due to reductions in early-life mortality), fertility rates have long been declining in richer and

[15] Of course, non-longtermist readers may see a bigger difference than longtermists do between 2270 and 2660!

better-educated countries. Fertility in France has been falling since the 1700s. Fertility in England and Wales has been falling since the 1800s. Fertility in Sweden, where records have long been of high quality, has never since matched its 1751 local peak, nor its 1823 local peak, nor its local 1901, 1944, nor 1990 peaks.[16] Fertility rates can fall for centuries and then stay low. We know that because they have.

More examples and more detail will require going beyond the ideal statistic (that is, going beyond completed cohort fertility) because not all countries collect adequate records and because completed cohort fertility is only available for cohorts old enough to be out of their childbearing ages. We cannot know how many children the women who were born in 2000 will have, on average, until at least about 2045 or 2050. But other measures of fertility exist and tell the same story. Figure 21.2 uses tempo-adjusted fertility to convert period fertility rates (how many children 20–24 year-olds, 25–29 year-olds, 30–34 year-olds, etc. are having today) to inform on cohort fertility rates (how many children these cohorts will have over their lifetimes).[17] The story is the same: Unreversing decline is found everywhere, even in countries with dissimilar societies, economies, and policies.

Most importantly, Figure 21.2 shows that two children per woman—the essential dividing line between population growth and decay—is no special stopping point as fertility rates decline. So far, every population that ever encountered the dashed line in Figure 21.2 just blew right through it. But of course they would. A population-level average of two is merely a theoretically interesting quantity in a demography textbook. A family can choose its own size—and may choose two (or zero or one or six)—but it is in nobody's power to choose two for the population average.

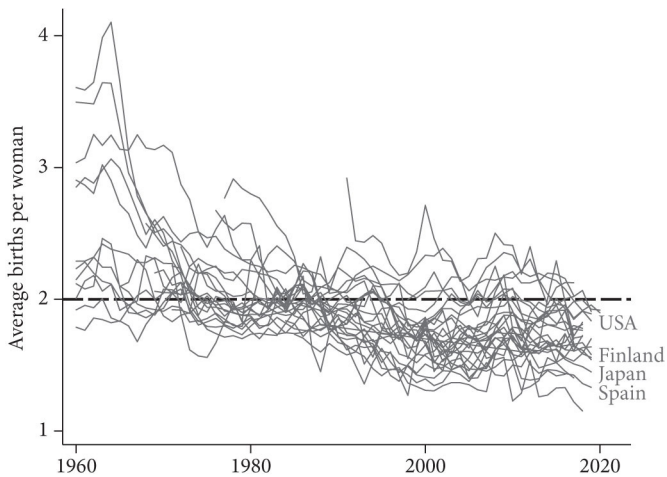


Figure 21.2 Tempo-adjusted total fertility does not stop at two: It keeps falling.
Source: Authors' drawing from the Human Fertility Database (2023).

[16] These years are from the Our World in Data graph 'Birth Rate, 1749 to 2020'; an article by Max Roser gives the source as 'International Historical Statistics' by Brian Mitchell. In the early 1990s, Sweden briefly climbed from its historic lows of the 1980s. During this baby boomlet, Swedish TFR momentarily kissed 2, and then promptly fell to a new historic low.

[17] Figure 21.2 is drawn using the Human Fertility Database (2023). It shows tempo-adjusted period fertility rates, for countries where fertility has fallen below 2. These are period rates, meaning descriptions of a point in time, rather than a cohort of women's observed fertility. Tempo adjustment incorporates the recognition that if births are being pushed to older ages, observing very few births among 20-year-olds today will underestimate total births over a lifetime. Tempo-adjusted period rates allow us to draw the graph farther into history (up to the present), but depend somewhat on the quality of the adjustment.

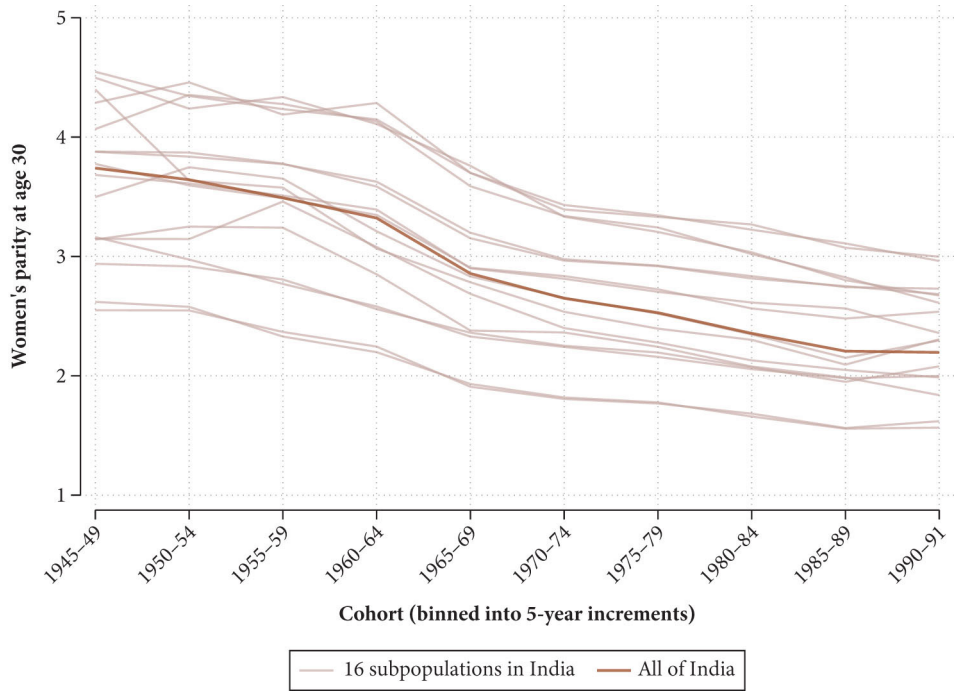


Figure 21.3 Fertility is falling for high and low fertility subpopulations: One example is 16 subpopulations of India.

Source: Authors' computations from the Demographic and Health Surveys for India (2021), updating a figure in Arenberg et al. (2022).

Figure 21.3 makes a similar point by focusing on India, a country that was the focus of 'overpopulation' rhetoric in the 20th century. It extends our previous work (Arenberg et al. 2022).[18] The vertical axis plots parity at age 30, an analogue of completed cohort fertility that computes the average number of children a birth cohort of women has had by age 30.[19] Each line is a demographically relevant group, split by education, geography, and religion. Figure 21.3 shows that: fertility is falling in every such Indian subpopulation; the gap between high and low fertility is narrowing; and several subpopulations are already below two children per woman.

Other places match the general pattern of long-run fertility decline. Fertility in China has never since been as high as in the mid-1960s. In the early 1990s, China's period total fertility rate fell below 2 and has never since exceeded it, according to UN statistics. It stands now, 30 years later, at an average of one and a quarter children per woman. Fertility for Latin America and the Caribbean, combined, has been falling for at least five decades, also in UN summary data, and is now below two. Sub-Saharan Africa is the only large region

[18] Figure 21.3 updates a graph that we first published with coauthors in Arenberg et al. (2022), here to include an additional, later round of survey data. The pattern is the same with the updated data as it was in our prior publication: All 16 lines slope downwards.

[19] This allows us to go later into history than completed cohort fertility would, because it lets us include birth cohorts that are only 30 years old at the time of the most recent survey. In India, childbearing tends to happen at young maternal ages and babies born to mothers over 30 are less common, so little is lost by using this measure here.

where average fertility is still above replacement,[20] but there, too, it has been falling for decades. The future of fertility in Africa is less certain than elsewhere, but improving education and declining mortality suggest that continued fertility decline is likely there, too (Kebede, Goujon, and Lutz 2019).

So the social scientific facts are consistent with continued fertility decline. We may not be able to be confident about all of the quantitative details of depopulation, but we saw in section 1 that those matter little on a longtermist timeline.[21]

At this point, you might be thinking of one or more objections. Perhaps you are thinking: Couldn't we just solve any problem with migration from high-fertility countries? Nope, because there's no migration on or off of Earth.[22] Or maybe you are thinking: Won't governments just sort this out if and when it becomes a problem? We'll explain in section 4 that no government ever has, nor does any government currently have any plausible plan for doing so.

Another common response is: Won't fertility rates equilibrate to 2, so that population size stabilizes? We reply: But why would that happen? There is no magical force to balance the number of births to the number of deaths. Right now, births exceed deaths. Soon, deaths will exceed births. No known equilibrating force will cause fertility rates to rise. None of the complex personal motivations and economic and cultural forces that drive individual decision-making generate a tendency to hold at 2 (as a population average). Ask the demographic experts of Japan, where TFR has been below replacement for 50 years, whether they've encountered evidence of this mysterious equilibrating force. So no, and Figure 21.2 has already shown that it hasn't happened in the countries that have so far experienced low fertility.

Another version of this is: Won't this just fix itself someday? This question has all of the hand-wavy dismissiveness of *Won't fertility equilibrate to 2?*, but doesn't mask its lack of a theoretical or empirical basis behind fancy words like equilibrating. We hope that this problem is fixed some day. But if it is, it would not be done without attention and investment.

Still another version is: Shouldn't technology solve this? Artificial wombs! robot nannies! AI tutors! longer lifespans and fertility-spans! We share the excitement for such a future, but we urge the techno-optimist to reflect that over the past two centuries—as lifespans have lengthened, as work days have shortened, as the world has gotten richer, as fertility has extended to later ages via technologies like in vitro fertilization, as care-work-supporting technology has proliferated (washing machines! disposable diapers! the Snoo!) —fertility has fallen, not risen. Technology expands opportunities, not family sizes, and this brand of

[20] Oceania is at 2.13, according to World Population Prospects figures for 2023 (United Nations 2022).

[21] And we can probably be more confident about demographic projections over the coming decades than you might think. In 1990, the UN projected that there would be 8.5 billion people in 2025. Now, 35 years later, it appears this will be off by only about 4%. The 1968 projection for 2000 was only wrong by 4.6%. (We wonder what other projections, predictions, or forecasts on other issues of concern to longtermists, could hope for such tested accuracy. Not many, we wager.) So why do different teams of demographers agree with each other, and how were those projections so accurate over a decades-long projection window? Population change turns out to be a simple dynamic system, in its biggest picture. Today's babies will grow up to be tomorrow's parents, but not for a few decades, so the progression of babies to adults is baked into population projections; the certainty that this fact creates is called 'population momentum'. This makes the nearer-term time path of population growth and decline amenable to precise forecasting. The key equation in determining the size of the population is the number of deaths each year and the number of births. As mortality rates continue to decline, they are bounded below by zero, so declining average mortality implies declining variance in the projection of mortality. The only major quantitative uncertainty that remains is the pace of the decline in fertility rates. Different population scientists disagree about this, but these disagreements do not make a huge difference at the scale of the Spike over the next century.

[22] We, the authors of this chapter, are all for freer migration. Migration might help some countries mitigate their near-term fiscal challenges. But this chapter is about *global* depopulation, so migration will not help.

fertility-oriented techno-optimism doesn't say why people will choose more children when their opportunities grow, instead of choosing something else.

But what, you might ask, about heritability (intergenerational transmission of high-fertility cultural practices)? Won't the Amish or some other high-fertility, perhaps religious, subpopulation expand indefinitely? For several reasons, no.[23] We have addressed this question at more length in Arenberg et al. (2022).[24] In the very long run (i.e., potentially after the coming few centuries of decline), two facts would have to be true for heritability to be a solution: First, fertility in a high-fertility subgroup would have to be high enough (certainly above 2, for example). We've already seen above that the 'high fertility' of high-fertility subgroups has been declining over the decades. High fertility used to mean six children per woman. Now it means 2.5. Before long, it may mean 1.8. Second, the children of high-fertility parents would have to be very likely to remain in their high-fertility cultural group. Where researchers have studied the empirical magnitude of these intergenerational correlations as they have played out in actual practice, they have found them to be positive, but small—too small, in fact, for the high fertility group to make much of a dent in overall population.[25] Culture notwithstanding, kids sometimes choose to do things different than their parents did. (If you have had a teenage child, you will not be surprised about what social scientists have documented in their studies.)[26]

Yes, it is theoretically possible that—against all historical precedence and against contemporary evidence and high quality forecasts—a sticky, high-fertility group identity emerges to stabilize the population. It is also theoretically possible (and simpler!) that, under an unprecedented social change, a new high-fertility norm could sweep the globe. Our claim is that it would be imprudent for longtermists to neglect the risk of sustained low fertility, given historical patterns and other well-documented social scientific facts.

3 Consequences

Why should a longtermist care about the scenarios like those consistent with what demographers expect to happen in the next 100 years? Our own study of long-term population projections is ultimately motivated by a population ethics that values the lives and

[23] Raftery and Ševčíková (2023: 95) cite Warren (2015) on the timeline of heritable fertility: 'For fertility, Warren (2015) has argued that a small subpopulation might become dominant over time if its members had consistently very high fertility, eventually leading to much higher than replacement fertility for the world population as a whole. His simulations showed, however, that it would require in the region of seven centuries for something like this to have a major global demographic impact, and its effect would likely still be modest in 2300, even if it started to happen immediately.'

[24] Arenberg et al. (2022) responds to arguments that applied population formulas from the mathematical biology literature to human demography. A fundamental reason why human fertility is different from, and less subject to tidy mathematics than, non-human animal population dynamics is that human fertility reflects intentional choices, technological change, culture, economic incentives, and other social influences. That is why we have this chapter to write, after all.

[25] Vogl (2020: 2976) summarizes: 'In populations with [total fertility rates] less than 3, differential fertility raises [the total fertility rate] by 4% on average.' This would not be enough to escape depopulation.

[26] Why, one might ask, would such a group remain cohesive and remain high-fertility, generation after generation, even as everyone else behaves differently? How would the social forces that keep a small social band in lockstep continue to discipline individual behavior when the group grows to no longer be a small band, but instead a group of hundreds of millions? And where, outside of their traditional geographies, would the growing group's members all live without changing their ways?

experiences of each person who might get to live a good life (rather than merely appreciating them instrumentally on the path to a longer-term future). We lay out our perspective in *After the Spike*. But we set that population ethics perspective of ours aside for this chapter.

Why might a small population in 2300, 2400, or 2500 threaten valuable and widespread flourishing in millennia thereafter? Our simple thesis, informed by population forecasts and economic theory and evidence, is this: Achieving the long and bright future that longtermists hope for might require a large number of people over the next few hundred years working to deliver it.

We mean the 'might' sincerely. Some optimistic longtermist might hope that intelligences not housed in human bodies will do the living and feeling that matters in the distant future. Fine. But should anyone be confident that it will happen in the next 300 years, before the world's number of scientists and engineers has spiraled downwards in a depopulation scenario? Or confident that progress towards these outcomes would not be slowed or halted entirely by a much smaller population in the near term? The timing matters. We may be racing towards a technologically enabled superabundance. But depopulation could win the race. We, the authors, are not confident of anything except that exponential decline in the number of births is the unavoidable mathematical consequence of fertility below 2.

If it does matter how many people live in the nearer term—whether because of population ethics, or economic growth, or environmental sustainability,[27] or extinction risk—then this would be what economists call an 'externality', meaning that no one individual, no one country, and no one generation has the right incentives (nevermind ability) to make the choices that would be best, all things considered.

Externalities are economists' classic cases where markets fail. Leaving an externality to run its course would make things worse, not better. Carbon emissions are a classic externality, so the solution is policy, public action, subsidized technological change, and coordination. If depopulation is indeed an externality that requires a coordinated, collective response, then there would be advantages to understanding the situation sooner rather than later. As Wolfgang Lutz and other demographers (Lutz, Skirbekk, and Testa 2006) have argued, once cultures, economies, and everyone's preferences organize around low fertility being normal, it may be hard to get out of a 'low fertility trap'.

What sort of externalities might a smaller population generate? Various mechanisms from macroeconomics suggest that a larger population would be likely to generate higher living standards, on average. This is not a fringe view in academic economics: Peters (2022) begins an abstract in a top journal with the summary that 'virtually all theories of economic growth predict a positive relationship between population size and productivity'.

[27] Calls to reduce the size of the human population are commonly heard, especially in popular media, as a suggested tool for decarbonization. Would depopulation be an effective tool of climate mitigation, that is, of reducing carbon emissions? No. Climate scientists have determined that humanity should seek to decarbonize in the next few decades. For example, the Biden White House has announced a strategy for the US to reach net-zero carbon emissions by 2050, which is 25 years from our writing. This is simply too soon for changing fertility rates to make much difference. Even if fertility rates changed significantly, the size of the population would maintain its trajectory for decades as today's stock of children and babies grow up into the coming decades' potential parents. A baby born yesterday will not have any children for decades (perhaps for 25 years or more), whatever fertility rates might be at that point. So fertility change is simply too slow to be a credible response to the urgency of decarbonization. Because of this, low fertility is not a constructive response to climate change. This argument was first made, to our knowledge, by Bradshaw and Brook (2014) and further developed by Budolfson and Spears (2021). For the detailed population and climate modeling supporting this argument quantitatively, see our paper with coauthors Kevin Kuruc, Sangita Vyas, and Mark Budolfson (Kuruc et al. 2022).

(Productivity here means on average, not in aggregate total.) If the totalist approach to population ethics (which says that more good lives is better) is correct, then the macroeconomics of scale effects tells us that we might reap those benefits for free—a larger population is better off on average and in total.[28]

A less productive economy with lower living standards may be less able to reach a flourishing future, or even to protect itself against certain types of risks. We review three economic mechanisms of scale effects from the literature: specialization, non-rival innovation, and fixed costs. Broadly, our point in this section is that population size matters for the economy, for living standards and for technological progress. Of course, other factors matter for economic outcomes, too.

Specialization and trade. Romer (1987) summarizes: 'The idea that specialization could lead to increasing returns is as old as economics as a discipline.' Specialization and trade is a core tool of modern economies. None of us reading this volume produces our own food, caretaking, medicine, electricity, clothes, transportation, software, or scholarly argumentation without any input from others. Specialization means that work is done by workers who know how to do it well. Specialization also prevents wastage and inefficiency in task-switching.[29]

You might be feeling confident that even a much smaller economy could continue to produce toasters and other consumer goods. But would it produce lifesaving new drugs, like the novel mRNA vaccine technology delivered just in time for combating COVID-19? No one can know. But we can know with high certainty that a world with orders of magnitude fewer people will probably have fewer molecular biologists. We can know that a smaller economy would be less complex, all else equal, and could not store and organize its human capital across so many humans, all else equal, which would limit our experts' ability to specialize.

Whatever technological breakthroughs you hope for that might usher in a new era of human flourishing, the economics of specialization tells us to not be confident that our shrinking world will produce it before we become too small to produce complex things.

Innovation and non-rival ideas. A second economic mechanism for scale effects is innovation. This mechanism is studied in the macroeconomics of endogenous economic growth, initially formalized by Romer (1986), developed further by Jones (2022), and a cornerstone of theories of humanity's escape from poverty (Kremer 1993; Galor and Weil 2000). The fundamental idea is that ideas, technologies, concepts, and strategies are non-rival. Non-rival is a term in economics which means that one person using a resource does not deplete the amount available for somebody else. Jones gives the example of the Pythagorean theorem: However often a builder consults the 3-4-5 triangle to build a right angle, the information remains fully intact, undepleted for somebody else to use.

The fact that ideas are non-rival generates scale effects because every person could potentially generate ideas that could then be used by everybody thereafter. On this view, it

[28] These economic mechanisms mean there is not an aggregate quantity-quality tradeoff: Despite all of the theoretical attention to the 'repugnant conclusion' of population ethics which trades off numbers of lives against quality of life, macroeconomists teach us that we should expect a larger population to be better off both on average and in total.

[29] As Paul Romer wrote, economists have recognized the importance of specialization in creating 'the wealth of nations' at least since Adam Smith published this observation under this title in 1776.

is no coincidence that humanity's huge expansion in technology coincided with its huge expansion in population. And if humanity depopulates, Jones has recently worked out in mathematical detail, technological progress and economic growth could end.[30]

Fixed costs, variety, and extinction risk. The final economic mechanism that might cause a small future to close off a bright future is fixed costs, which are featured in trade and geographic economics (Krugman 1991). The costs of economic activities are divided in microeconomic theory into two costs: variable costs, which scale with the quantity produced, and fixed costs, which do not. Product differentiation and fixed costs are important reasons why real economies are not the perfectly competitive economies of introductory textbooks. In particular, fixed costs can be a barrier to market entry for a new firm, if there will not be enough customers to cover the fixed costs of the new business.

Fixed costs do not merely apply to businesses. Consider greenhouse gasses. Imagine the year after humanity reaches the point of net-zero annual emissions. Going forward from that point, there will be a fixed amount of greenhouse gasses in the atmosphere, accumulated since the start of the industrial era. Intentional 'negative emissions' technology, such as sequestering carbon underground, could reduce that fixed stock of greenhouse gasses at an economic cost. But only if somebody chooses to pay it. A larger total economy, which could be achieved by having a more populous total economy, would have more resources that it could choose to devote to the fixed cost of negative emissions.

Some existential threats take such a fixed-cost form.[31] They are exogenous to human population size and would arrive at a historical time that is independent of human activity, such as, perhaps, a large asteroid heading towards Earth. Consider the following simple model of such a situation, which we introduce in *After the Spike*: An exogenous threat has arisen which will kill all humans (however many that may be) unless a large cost is paid to avoid it within a certain time period.

This large cost is fixed, in the microeconomic sense, because it must be paid to deflect the asteroid, whether 'killing all of us' means killing 10 billion or killing 10 million. The cost of avoiding the disaster does not scale with population size. And here is where population scale effects come in. Which do you think would be more likely to successfully pay the fixed cost in time: the larger population and economy or the smaller one? Probably the larger one: After all, if both societies were equally rich per capita (and they wouldn't be; we're handicapping the larger population here, relative to what the macroeconomic-growth literature teaches us), then the larger society could out-produce the smaller society in rockets or anything else needed to avoid disaster. In the larger economy (10 billion of us), a 0.1% tax levied to fund the response effort would outstrip a 50.0% tax levied in the smaller economy (10 million of us). At least some existential risks would be more likely to be survived by a larger population.

There are several reasons to stop neglecting depopulation risk now, rather than waiting a few decades. First, as we discuss in the next section, humanity presently has no policy or

[30] Indeed, depopulation could be even worse than in Jones's model because Jones does not incorporate depreciation of technology (Eden and Kuruc 2022). This means that Jones does not include a cost of keeping ideas and knowledge in existence and available in usable form. Maybe such knowledge-depreciation would happen slowly enough that rapid improvements in information technology would more than make up for the losses due to depopulation. Or, maybe not.

[31] Other threats may be less likely or more likely to arise if more people are alive at a time; we ignore those threats here.

technology that could reverse the decline if that turned out to be desirable. It's time to make progress on understanding possible responses so there is an option of action. Depopulation could be very fast (counting in centuries), and there might not be much time to course-correct, if it turns out to be a bigger problem than you think. Second, it might be hard to ever convince most people of the positive externalities of scale—that the optimal number of people to have around is more than whatever they are used to believing. So decline might be a one-way ratchet. Perhaps the best humanity can ever do is stabilize. If so, stabilizing soon, at a higher level, means many more stick figures in humanity's future.

4 Responses

To know how and whether to respond to depopulation, we need to know two things: first, why so many people in such different societies are choosing to have fewer children, and second, what responses might be available, given the reality of human societies and their governments. We offer the unsatisfying—but scientifically grounded—observation that nobody yet has an answer to either question. So the most important response is to first invest in learning more.

Why is sustained fertility decline seen everywhere? The available claims and evidence do not point toward a clear answer. There is a range of theories, roughly divisible into economic theories and cultural theories. Economic theories of fertility decline emphasize that, in richer societies, children change from being a source of economic support for their parents to being an expensive consumption good for their parents. Or that the calculus changes to favor investing more each in a smaller number of children—especially so in economies where there are large benefits to be gained from many years of expensive investments in education and human capital. Cultural theories point, instead, to a change in values, away from family or traditional roles as a motivating factor in life and towards adults' own fulfillment or enjoyment (Lesthaeghe 2010).[32]

There is something to learn from each of these families of theories. But no theory is yet widely accepted, and no theory fits all of the facts (Doepke et al. 2022). For example, the economic theory of the quantity-quality tradeoff, where parents have fewer children in order to invest more in the education of each one, cannot immediately explain societies where many people choose to have *no* children. Similarly, differences across US states in the economic costs of children do not explain differences in the pace of fertility decline (Kearney, Levine, and Pardue 2022).

Most importantly, fertility decline is a convergent phenomenon, happening in many places. Theories of female paid labor force participation as a binding constraint cannot explain India, especially south India, where fertility rates are below replacement even though only a minority of women work in the paid labor force (Gietel-Basten, Spears, and Visaria 2022).

[32] A more specific version of this is the 'incomplete gender revolution' theory, according to which the driving force is the fact that women are increasingly both free to and, in some cases, economically compelled to pursue education and paid labor market work, but do not receive support from partners and other family members in the work of parenting (Esping-Andersen 2009).

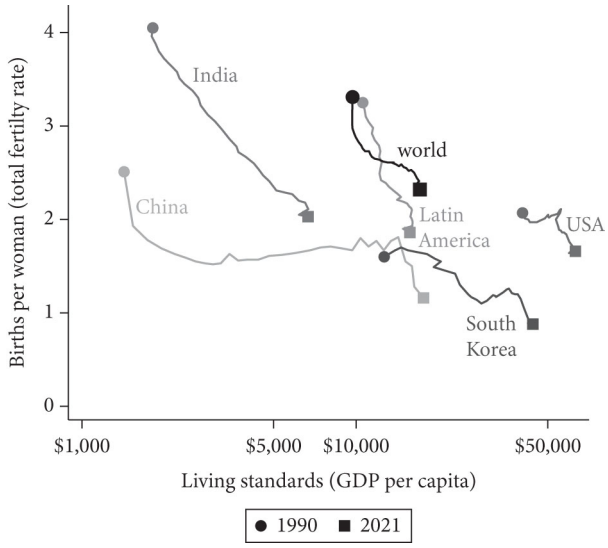


Figure 21.4 Living standards have expanded as fertility has declined.

Note: Our computations with data from UN World Population Prospects fertility data and World Bank ICP GDP data.

Figure 21.4 summarizes what we see as one of the most important facts: Because declining fertility has accompanied rapid global economic growth, average living standards have expanded radically over the same decades in which average fertility rates have been falling. That should teach us to be skeptical of overly simple economic theories that children are less affordable than they used to be: Although our point is not that anyone *should*, families *could* choose a 1990 consumption bundle and a 1990 fertility level and still have more money left over and better things (better devices, better houses, better health care, safer cars . . .) than people in 1990 did.

The second question to ask about responding to low fertility is what policies and programs might be feasible and might be good ideas. Here, too, there is much to learn. The idea that social policy can meaningfully move fertility has a hard time fitting the facts of international comparisons.

Welfare states are larger in Europe than in the United States, but fertility is lower in Europe, on average. Rightly or wrongly, US progressives hold up Sweden as a model of what pro-parent policy-making could be. But the average woman in Sweden in 2018 had children at a birth rate of 1.76, compared to 1.73 in the US. In 2019, both countries fell to 1.70, by chance matching Denmark. Norway and Finland, in case you don't trust these examples, dropped in 2019 to 1.53 and 1.35, respectively.[33] These four European countries each spend twice as much as the US spends on family benefits, as a fraction of gross domestic product (GDP). The sort of US$3,600 child tax credit that was briefly implemented, debated, and then eliminated in US politics over the past few years is small relative to benefits in these countries, is small relative to the costs of parenting an extra child, and is unlikely to make

[33] These period total fertility rates are from the World Bank World Development Indicators.

much of a difference in aggregate fertility outcomes. Whether a much larger benefit would make a difference is an unsettled social scientific question that is amenable to experimentation and other forms of research.[34] But we don't have that answer yet.[35]

The lessons from experience with fertility policy are both strong and negative. Governments in some times and places have tried to compel people to have children they don't want to have; governments in some times and places have tried to compel people not to have children they want to have. These policies have done terrible harm to people's lives without changing aggregate fertility rates. If you would like to learn more, read Hartmann (1987) or Connelly (2010).

It is not within the scope of this essay to adequately review the social science and history of population policy. But because we are often confronted with people who seem to believe that governments can choose fertility rates as easily as the Federal Reserve Bank chooses interest rates,[36] we'll merely note that the widely cited one-child policy from China is widely misunderstood. In particular, birth rates fell dramatically over the decade *prior to* the policy (there was a different fertility policy at that time, too). Yes, China's population policy was a harmful repression of the freedom of individuals to choose their lives. And, in the other direction, anti-abortion policy elsewhere has been and is harmful and repressive, too. But, as Susan Greenhalgh (2018) and others have documented, China was also experiencing large socioeconomic changes of the sort that are known to have contributed to fertility decline elsewhere.

So what then should we do? The next step is to learn more about low fertility and depopulation: the causes, the consequences, and the possible responses. This requires investment in basic science: improved population forecasts and scenarios over a longer horizon; research to better understand the causes of global fertility decline; medical and pharmaceutical research to make family planning (that is, achieving all and only wanted children) surer, safer, and more convenient; experimentation and other research of the potential impacts of child tax credits, generous parental leave mandates, free daycare, and other support for parents, families, and caregivers of all kinds; and so on.

Any longtermist who is disappointed that we do not have shovel-ready solutions to depopulation should reflect that many experts who are worried about unaligned artificial general intelligence believe that more basic research is needed there too, before we have solutions to implement.

Climate research used to be in just this state as well, and offers lessons. Depopulation is projected to begin in about 60 years (Lutz et al. 2014; IHME 2020; United Nations 2022). About 60 years ago, in President Lyndon Johnson's administration in the US, the White House recognized that carbon dioxide is an important pollutant that would

[34] For example, a philanthropic program to randomize large cash transfers to families with the aim of measuring the fertility response over the following decade could provide needed evidence. A state government might also randomize a large benefit—for example, as Oregon did in awarding a health insurance benefit via lottery in 2008.

[35] Here is another high-level way to see that we have no tried-and-tested, ready responses: Consider the many European and East Asian countries with much-publicized pro-natalist policies. Consider also that these countries have retained enduring low fertility rates (Gietel-Basten 2019).

[36] Whether population policy can actually succeed at achieving sustained changes to fertility is, contrary to popular belief, hardly settled in favor of the population controllers. As Babiarz et al. (2018) explain in their examination of China's Later Longer Fewer campaign: 'The extent to which population policy is able to influence reproductive behavior in developing countries has been fiercely debated in economics and demography.'

change the climate by changing the atmosphere. We, now, do not know what to do about depopulation, just as the best climate scientists of Johnson's time could not have directed today's energy policies. But they knew some big facts—in particular, that emitting carbon dioxide would eventually warm the Earth. And we know some big facts—in particular, that sustained below-replacement fertility would depopulate the Earth at an exponential pace.

Climate policy is achieving more, today, than it otherwise would be if the research and debates of the past 60 years had not gotten started, even without having all the answers. The relatively small investment in climate research a half century ago is paying massive dividends now. Of course, it would have been even better if they had done more! Following their example, longtermists and others should turn more of their attention and research to depopulation. We will never achieve shovel-ready projects to address this challenge if we do not begin the work of basic scientific research to understand this problem and potential responses. And it is likely there will be shovel-ready projects to be found. A $3,600 child tax credit doesn't move the needle. But would even more ambitious public support for parents, if it could be given in a way that empowers with freedom and consent? More years of paid parental leave, if it could be given in a way that advances, rather than threatens, gender equality at work? The space of unevaluated, economically and technologically feasible responses is huge. Basic research could guide us towards which to invest in evaluating first.

5 Conclusion

Because this brief chapter is intended to invite longtermists and population scientists into dialogue, it did not have enough to say about some important issues. It did not have enough to say about gender inequality. Up to now and in any foreseen future, all children are birthed by pregnant females.[37] This fact alone makes reproduction unequal, without even considering the gender inequality that our societies and economies layer on top of what biology has endowed us with. No understanding of or response to low fertility can ignore gender inequality. If society does not share the burden of producing the next generations, lifting the burden on mothers and other caretakers, then we should not be surprised if they do not take it up. If a flourishing next generation is a public good, an economist might say, too many men have been free riding on women's contributions.[38]

This chapter did not have enough to say about population ethics. One ethical foundation of any response to depopulation is to insist on reproductive freedom: Any good future is one where whether to parent or not is a free choice. Another fact that raises questions of values and ethics is that billions of good lives that might have been lived over the next few centuries will not be if we do not begin work to address depopulation. Even people who do not count themselves as longtermists could find this an important reason to make better

[37] There is a tension between, on the one hand, recognizing the gender inequality that puts so much of the burden of care work and parenting on women and, on the other hand, celebrating the growing freedom in gender identities, which recognizes that not all pregnant people who bear that burden identify as *women*. We refer the reader to Foster's (2020) discussion of this tension in *The Turnaway Study* and follow her use of 'women'.

[38] Perhaps if more of the powerful men of the past spent many nights exhausted, soothing an upset baby who couldn't quite figure out how to eat or sleep, we would have already made more progress.

understanding of depopulation a research priority—just like, for example, animal welfare and global health are also important priorities.

There may be many future people. Or there may be strikingly few. We hope that the world begins investing serious attention to depopulation so that someone can someday know what to do about it. We are skeptical that anyone yet knows exactly what to do about it. And there are risks of reckless action.[39]

We do not know what is going to happen. One possibility is an unprecedented reversal in fertility trends, perhaps bolstered by unprecedented policy investments in children, gender equality, care work, and the freedom to parent or not to parent. Another possibility is that humanity slides down the Spike. It is easy to imagine that, instead of supporting one another, future people comfort themselves with art and culture and stories to tell one another that depopulation is good.

Perhaps we risk political naïveté to hope that politics can construct a response to depopulation that expands people's freedom and options, respects their autonomy, and better supports those who want to have more children (and raise them safely and healthily) while also supporting those who want not to—all while balancing the many other worthy demands on our politics and policy. It would not do the causes supported by longtermists (or any other good cause) any favors to ignore the political risks and history. But in the face of the Spike, we do hope.

Acknowledgments

We are grateful for comments on drafts of this chapter from Gustav Alexandrie, Kathleen Broussard, Mark Budolfson, Diane Coffey, Maya Eden, Bob Fischer, Lauren Gaydosh, Aashish Gupta, Johan Gustafsson, Payal Hathi, Kevin Kuruc, Ester Lazzari, Melissa LoPalo, Christian Tarsney, Sangita Vyas, Gage Weston, and Bridget Williams, and from the students of Professor Gaydosh's graduate demography seminar. This published and peer-reviewed version supersedes the prior draft version that was circulated for peer review and comment under a different title.

References

Arenberg, S., Kuruc, K., Franz, N., Vyas, S., Lawson, N., LoPalo, M., Budolfson, M., Geruso, M., and Spears, D. (2022), 'Research Note: Intergenerational Transmission Is Not Sufficient for Positive Long-Term Population Growth', in *Demography* 59/6: 2003–2012.

Basten, S., Lutz, W., and Scherbov, S. (2013), 'Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility', in *Demographic Research* 28: 1145–1166.

Bradshaw, C. J. and Brook, B. W. (2014), 'Human Population Reduction is Not a Quick Fix for Environmental Problems', in *Proceedings of the National Academy of Sciences* 111/46: 16610–16615.

[39] When we shared the ideas behind this chapter with a public health colleague, she warned us: 'But you know what is going to happen', meaning that the political proponents of social inequality and coercion will use the threat of depopulation as motivation for another round of repressive politics (using arguments like ours as cover). This concern comes from a place of wisdom and deserves our attention and respect.

Babiarz, K. S., Ma, P., Miller, G., and Song, S. (2018), 'The Limits (and Human Costs) of Population Policy: Fertility Decline and Sex Selection in China under Mao', National Bureau of Economic Research working paper.

Budolfson, M. and Spears, D. (2021), 'Population Ethics and the Prospects for Fertility Policy as Climate Mitigation Policy', in *Journal of Development Studies* 57/9: 1499–1510.

Connelly, M. (2010), *Fatal Misconception: The Struggle to Control World Population* (Harvard University Press).

Demographic and Health Surveys. (2021), https://dhsprogram.com/

Doepke, M., Hannusch, A., Kindermann, F., and Tertilt, M. (2022), 'The Economics of Fertility: A New Era', National Bureau of Economic Research Working Paper w29948.

Eden, M. and Kuruc, K. (2022), 'Marginal Benefits of Population: Evidence from a Malthusian Semi-Endogenous Growth Model', Population Wellbeing Initiative Working Paper 2305.

Esping-Andersen, G. (2009), *Incomplete Revolution: Adapting Welfare States to Women's New Roles* (Polity).

Foster, D. G. (2020), *The Turnaway Study* (Scribner).

Galor, O. and Weil, D. N. (2000), 'Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond', in *American Economic Review* 90/4: 806–828.

Gietel-Basten, S. A., Sobotka, T., and Zeman, K. (2014), 'Future Fertility in Low Fertility Countries', in W. Lutz, W. P. Butz, and K. C. Samir (eds.), *World Population and Human Capital in the Twenty-First Century* (Oxford University Press), 39–146.

Gietel-Basten, S. (2019), *The 'Population Problem' in Pacific Asia* (Oxford University Press).

Gietel-Basten, S. A., Spears, D., and Visaria, L. (2022), 'Low Fertility with Low Female Labor Force Participation in South India', in *PAA 2022 Annual Meeting* (Population Association of America).

Greenhalgh, S. (2018), 'Making Demography Astonishing: Lessons in the Politics of Population Science', in *Demography* 55/2, 721–731.

Hartmann (1987), *Reproductive Rights and Wrongs: The Global Politics of Population Control* (Harper & Row).

Human Fertility Database. (2023), https://www.humanfertility.org/

Institute for Health Metrics and Evaluation (IHME). (2020), 'Population Forecasting'. https://vizhub.healthdata.org/population-forecast/

Jones, C.I. (2022), 'The End of Economic Growth? Unintended Consequences of a Declining Population', in *American Economic Review* 112/11: 3489–3527.

Kaneda, T. and Haub C. (2022), 'How Many People Have Ever Lived on Earth?', in *Population Reference Bureau*, https://www.prb.org/articles/how-many-people-have-ever-lived-on-earth/

Kearney, M.S., Levine, P.B., and Pardue, L. (2022), 'The Puzzle of Falling US Birth Rates since the Great Recession', in *Journal of Economic Perspectives* 36/1: 151–176.

Kebede, E., Goujon, A., and Lutz, W. (2019), 'Stalls in Africa's Fertility Decline Partly Result from Disruptions in Female Education', in *Proceedings of the National Academy of Sciences* 116/8: 2891–2896.

Kremer, M. (1993), 'Population Growth and Technological Change: One million BC to 1990', in *Quarterly Journal of Economics* 108/3: 681–716.

Kuruc, K., Vyas, S., Budolfson, M., Geruso, M., and Spears, D. (2022), 'Is Less Really More? Comparing the Climate and Productivity Impacts of a Shrinking Population', Population Wellbeing Initiative Working Paper w2302.

Krugman, P. (1991), 'Increasing Returns and Economic Geography', in *Journal of Political Economy* 99/3: 483–499.

Lesthaeghe, R. (2010), 'The Unfolding Story of the Second Demographic Transition', *Population and Development Review* 36/2: 211–251.

Lutz, W., Butz, W. P., and Samir KC (2014), *World Population and Human Capital in the Twenty-First Century* (Oxford University Press).

Lutz, W., Skirbekk, V., and Testa, M. R. (2006), 'The Low-Fertility Trap Hypothesis: Forces that May Lead to Further Postponement and Fewer Births in Europe', in *Vienna Yearbook of Population Research*: 167–192.

MacAskill, W. (2022), *What We Owe the Future* (Basic Books).

Newberry, T. (2021), 'How Many Lives Does the Future Hold?' Global Priorities Institute Technical Report T2-2021, https://globalprioritiesinstitute.org/how-many-lives-does-the-future-hold-toby-newberry-future-of-humanity-institute-university-of-oxford/

Ord, T. (2020), *The Precipice* (Hachette Books).

Peters, M. (2022), 'Market Size and Spatial Growth—Evidence from Germany's Post-War Population Expulsions', *Econometrica* 90/5: 2357–2396.

Raftery, A. E., and Ševčíková, H. (2023), 'Probabilistic Population Forecasting: Short to Very Long-Term', in *International Journal of Forecasting* 39/1: 73–97.

Romer, P. M. (1986), 'Increasing Returns and Long-Run Growth', in *Journal of Political Economy* 94/5: 1002–1037.
Romer, P. M. (1987), 'Growth Based on Increasing Returns Due to Specialization', in *American Economic Review* 77/2: 56–62.
Roser, M. (2022), 'The Future Is Vast: Longtermism's Perspective on Humanity's Past, Present and Future", in *Our World in Data*, https://ourworldindata.org/longtermism
Spears, D., Vyas, S., Weston, G., and Geruso, M. (2024), 'Long-Term Population Projections: Scenarios of Low or Rebounding Fertility', *Plos One*: 1–16.
Spears, D. and Geruso M. (2025), *After the Spike: Population, Progress, and the Case for People* (Simon and Schuster).
United Nations World Population Prospects. (2022), https://population.un.org/wpp
Vogl, T. S. (2020), 'Intergenerational Associations and the Fertility Transition', in *Journal of the European Economic Association* 18/6: 2972–3005.
Warren, S. G. (2015), 'Can Human Populations Be Stabilized?', in *Earth's Future*, 3/2: 82–94.

22

Existential Risk from Power-Seeking AI

Joe Carlsmith

1 Introduction

Some worry that misaligned artificial intelligence (AI) will pose an existential risk to humanity.[1] In this essay, I formulate and examine what I see as the core argument for such a concern. In brief, and put roughly, the argument is that by 2070:[2]

1. It will become possible and financially feasible to build relevantly powerful and agentic AI systems.[3]
2. There will be strong incentives to do so, conditional on (1).
3. It will be much harder to build aligned (and relevantly powerful/agentic) AI systems than to build misaligned (and relevantly powerful/agentic) AI systems that are still superficially attractive to deploy, conditional on (1) and (2).
4. Some such misaligned systems will seek power over humans in high-impact ways, conditional on (1)–(3).
5. This problem will scale to the full disempowerment of humanity, conditional on (1)–(4).
6. Such disempowerment will constitute an existential catastrophe, conditional on (1)–(5).

These claims are extremely important if true. My aim is to investigate them.[4]

My current view is that there is a disturbingly substantive chance (i.e. greater than 10%) that all of these claims are true, and that many people alive today—including myself—live to see humanity permanently disempowered by AI systems we've lost control over.[5] That is: I

[1] For classic arguments and other resources, see e.g. Yudkowsky (2008), Bostrom (2014), Tegmark (2017), Christiano (2019), Russell (2019), Ord (2020), Ngo (2020), Karnofsky's (2021; 2022), Ngo, Chan, and Mindermann (2022), and Cotra (2021). By 'existential risk', I mean a risk that threatens to destroy humanity's longterm potential (here I am following Ord (2020: 27)).

[2] I'm focusing on 2070 because I want to keep vividly in mind that I and many other readers (and/or their children) should expect to live to see the claims at stake here falsified or confirmed. That said, the main arguments don't actually require the development of relevant systems within any particular period of time (though timelines in this respect can matter to e.g. the amount of evidence that present-day systems and conditions provide about future risks).

[3] I define various of the terms in this argument—e.g. 'relevantly powerful and agentic', 'misaligned'—more precisely below.

[4] For somewhat complicated reasons, the sections below do not correspond perfectly to the argument's premises.

[5] In a longer report on which this essay is based (see Carlsmith (2022)), I try to get more quantitative purchase on the level of risk, by assigning probabilities to the various premises in the argument, but I won't do so here. A striking fraction of experts in the field, however, seem likely to agree that the risk is at least 10%: in Stein-Perlman, Grace, and Weinstein-Raun's (2022) survey of more than 700 AI researchers who had recently published at NeurIPS or ICML (major machine learning conferences), 48% of respondents gave at least 10% chance that the long-run effect of AI will be 'extremely bad (e.g. human extinction)'. The median respondent said 5%.

Joe Carlsmith, *Existential Risk from Power-Seeking AI* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Joe Carlsmith 2025. DOI: 10.1093/9780191979972.003.0022

view this as a problem of grave importance. My hope, here, is to facilitate productive debate about it.

2 Backdrop

The specific arguments I'll discuss emerge from a broader backdrop picture, which I'll gloss as:

1. Intelligent agency is an extremely powerful force for controlling and transforming the world.
2. Building agents much more intelligent than humans is playing with fire.

I'll start by briefly describing this picture, as it sets an important stage for the discussion that follows.

Of all the species that have lived on the Earth, humans are clearly strange. In particular, we exert an unprecedented scale and sophistication of intentional control over our environment. Consider, for example, the city of Tokyo, or the Large Hadron Collider, or a large coal mine.

What makes this possible? Something about our minds seems centrally important. We can plan, learn, communicate, deduce, remember, explain, imagine, experiment, and cooperate in ways that other species can't. These cognitive abilities—employed in the context of the culture and technology we inherit and create—give us the power, collectively, to transform the world. Let's call this loose cluster of abilities 'intelligence', though very little will rest on the term.

And humans aren't just smart: we're also *agentic*. That is (loosely): we pursue objectives, guided by models of the world. We have cities, particle accelerators, and coal mines because we were *trying* to build them.

It seems possible, in principle, to build agentic cognitive systems—both biological and artificial—whose intelligence significantly exceeds our own. And in the context of artificial agents, the differences between brains and computers—in possible speed, size, available energy, memory capacity, component reliability, input/output bandwidth, and so forth—make the eventual possibility of very dramatic differences in ability especially salient.

But the choice to build such superhumanly intelligent artificial agents should be approached with extreme caution.[6] As humanity's impact on the Earth illustrates, intelligent

[6] Some articulate this view by appeal to the dominant position of humans on this planet, relative to other species (see e.g. Bostrom (2015) and Russell (2019: ch. 5) on the 'Gorilla Problem', Ord (2020) and Ngo (2020) on the 'second species argument'). For example: some argue that the fate of the chimpanzees is currently in human hands, and that this difference in power is primarily attributable to differences in intelligence, rather than e.g. physical strength. Just as chimpanzees—given the choice and power—should be careful about building humans, then, we should be careful about building agents more intelligent than us. This argument is suggestive, but far from airtight. Chimpanzees, for example, are themselves much more intelligent than mice, but the 'fate of the mice' was never 'in the hands' of the chimpanzees. What's more, the control that humans can exert over the fate of other species on this planet still has limits, and we can debate whether 'intelligence', even in the context of accumulating culture and technology, is the best way of explaining what control we have. More importantly, though: humans arose through an evolutionary process that chimpanzees did nothing to intentionally steer. Humans, though, will be able to control many aspects of processes we use to build and empower new intelligent agents.

agency is a force of formidable potency. If we unleash much more of this force into the world, via new, more intelligent forms of non-human agency, it seems reasonable to expect dramatic impacts, and reasonable to wonder how well we will be able to control the results.

I'll focus on a particular version of this worry, centered on the following hypothesis: that by default, suitably strategic and intelligent agents, engaging in suitable types of planning, will have instrumental incentives to gain and maintain various types of power (call this 'power-seeking'), since this power will help them pursue their objectives more effectively. The worry is that if we create and lose control of such agents, the result won't just be *damage* of the type that occurs when a plane crashes, or a nuclear plant melts down—damage which remains passive, for all its costs. Rather, the result will be highly capable, non-human agents actively working to gain and maintain power over their environment—agents in an *adversarial* relationship with humans who don't want them to succeed.

Nuclear contamination is hard to clean up, and hard to stop from spreading. But it isn't *trying* to spread—and certainly not with greater intelligence than the humans trying to contain it. But the power-seeking agents just described *would* be trying, in sophisticated ways, to undermine our efforts to stop them. If such agents are sufficiently capable, and/or if sufficiently many of such failures occur, humans could end up permanently disempowered.

In principle, then, sophisticated AI agents with problematic goals could represent an unprecedented threat to the human species. Let's look more closely at whether to expect this threat to arise in practice.

3 APS systems

This section discusses in more detail the type of AI systems I'm worried about, and the timelines to their development.

I'll focus on AI systems with three key properties.

1. *Advanced capability*: they outperform the best humans on some set of tasks which when performed at advanced levels grant significant power in today's world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation).[7]

The aim here is to hone in on systems whose capabilities make any power-seeking behavior they engage in worth taking seriously (in aggregate) as a potential route to the disempowerment of roughly all humans. Such a condition does not, I think, require meeting various stronger conditions sometimes discussed[8]—for example, 'human-level

[7] An AI system with these capabilities can consist of many smaller systems interacting, but it should suffice to ~fully automate the tasks in question.

[8] A level of AI progress that disempowered all humans would constitute 'transformative AI' in the sense used by Karnofsky (2016): e.g. 'AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution'. But *that* sort of transformation is precisely the type of thing we're trying to forecast; e.g. it's the result, not the cause. And disempowerment does not require other, more mechanistic standards for transformation—e.g. economic growth proceeding at particular rates (though we can argue, here, about the economic value that AI progress sufficient to disempower humans would represent).

AI',[9] 'superintelligence',[10] or 'artificial general intelligence (AGI)'.[11] That said, I'm erring, here, on the side of including 'weaker' systems—including some that might not, on their own (or even in aggregate), be all that threatening.[12]

2. *Agentic planning:* they make and execute plans, in pursuit of objectives, on the basis of models of the world.

The aim here (and with the next property, 'Strategic awareness') is to hone in on the type of goal-oriented cognition required for arguments about the instrumental value of gaining/maintaining power to be relevant to a system's behavior.[13] That is, in order to have instrumental incentives to seek power, a system needs to have objectives that power-seeking promotes, and its behavior needs to be sensitive to the incentives those objectives create.

We can argue about exactly what is required for concepts like 'planning', 'pursuing objectives', and 'using models of the world' to apply—and indeed, muddiness around abstractions in this vicinity seem to me a key way that thinking on this topic (including my own) might go astray.[14] I take it, though, that humans do these things, and that AI systems can, too. I'm talking about the ones that do (or at least, that do something close enough to justify predicting their behavior on this basis).

3. *Strategic awareness:* the models they use in making plans represent with reasonable accuracy the causal upshot of gaining and maintaining power over humans and the real-world environment.

The aim here is to hone in on the type of world-modeling required to notice and respond to incentives to seek power, where they exist. Clearly, this capability comes in degrees. But broadly and loosely, we can think of a strategically aware, planning agent as possessing models of the world that would allow it to answer questions like 'what would happen if I had access to more computing power' and 'what would happen if I tried to stop humans from turning me off' about as well as humans can (and using those same models in generating plans).

[9] This is used in various ways, to mean something like (i) a single AI system that is in some generic sense 'as intelligent' as a human; (ii) a single AI system that can do anything that a given human (an average human? the 'best' human? any human?) can do; (iii) a level of automation such that unaided machines can perform roughly any task better and more cheaply than human workers (see Grace et al. (2018)). My favorite is (iii).

[10] Bostrom (2014: ch. 2) defines this as 'any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest'. A related concept requires cognitive performance that in some loose sense exceeds all of human civilization—though exactly how to understand this isn't quite clear (and human civilization's 'cognitive abilities' change over time).

[11] 'AGI' is sometimes used as a substitute for some concept of 'human-level AI'; in other contexts, it refers specifically to some concept of human *learning* ability (see e.g. Selsam (2018) and Yudkowsky (n.d.b.) or some method of creating systems that can perform certain tasks (see Ngo (2020) on 'generalization-based approaches'). 'AGI' is often contrasted with 'narrow AI'—though in my opinion, this contrast too easily runs together a system's ability to *learn* tasks with its ability to *perform* them. And sometimes, a given use of 'AGI' just means something like 'you know, the big AI thing; *real* AI; the special sauce; the thing everyone else is talking about'.

[12] Put another way, I'm erring on the side of 'necessary' rather than 'sufficient' for riskiness. I have yet to hear a good capability threshold that is both necessary *and* sufficient—and I'm skeptical that one exists.

[13] We can imagine cases in which AI agents end up valuing power for its own sake, but I'm not going to focus on those here.

[14] See the longer report, cited in footnote 5, for more discussion.

Let's call a system with all three of these properties an Advanced, Planning, Strategically aware system (or 'APS system').

Will it become possible and financially feasible to build APS systems before 2070? I think that this is more likely than not.[15] However, I won't attempt to examine the issue here.

Perhaps the right probability here is lower. But I doubt that it is far lower. Less than 10%, for example, seems to me unreasonable (and I don't think that the difficulty of forecasts like these licenses assuming that the probability here is very low, or treating it that way implicitly).

4 Incentives

Let's assume, then, that it will become possible and financially feasible to develop APS systems. Should we expect relevant actors to do so, especially on a widespread scale?

It seems likely that there will be strong economic and political incentives to automate advanced capabilities. But building strategically aware agentic planners might not be the only way to do this. After all, many tasks that one might wish to automate—translating languages, classifying proteins, predicting human responses, etc.—don't seem to require agentic planning or strategic awareness (at least at current levels of performance).

Nevertheless, I think, there are strong reasons to expect that AI progress will push in the direction of APS systems. I will focus on three.

The first and strongest reason is that agentic planning and strategic awareness seem quite *useful*. Agentic planning is a very powerful and general way of interacting with an environment—especially a complex and novel environment that doesn't afford much room for trial and error—in order to produce favored outcomes. Many tasks humans care about (creating and selling profitable products, designing and performing successful scientific experiments, achieving social and political goals) have this structure. So do many of the sub-tasks involved in those tasks (e.g. efficiently gathering and synthesizing relevant information, communicating with stakeholders, managing resource allocation, etc.). And even when agentic planning isn't strictly required to achieve a valuable outcome, it will often help. So if our AI systems can't engage in agentic planning, then the scope of what they can do seems, naively, like it will be severely restricted.

So, too, with strategic awareness. Strategic awareness is closely connected to a basic capacity to 'understand what is going on', interact with other agents (including human agents) in the real world, and recognize available routes to achieving objectives—all of which seem very useful to performing tasks of the type just described. Indeed, to the extent that humans care about the strategic pursuit of e.g. business, military, and political objectives, and want to use AI systems in these domains, it seems like there will be incentives to create AI systems with the types of world models necessary for very sophisticated and complex types of strategic planning—including planning that involves recognizing and using available levers of real-world power.

[15] This view emerges partly from a series of investigations at Open Philanthropy into AI timelines, summarized in Karnofsky (2021); partly from public forecasts; partly from eye-balling recent progress in deep learning myself; and partly from deference to some of Open Philanthropy's technical advisors.

Here is a second reason to expect APS systems. Even if some task doesn't *require* agentic planning or strategic awareness, it may be that creating APS systems is the only route, or the most efficient route, to automating that task, given available techniques. For example: perhaps the best way to automate a wide range of tasks is to create AI systems that can learn new tasks with very little data.[16] And we can imagine scenarios in which the best way to do *that* is by training agentic planners with high-level, broad-scope pictures of 'how the world works', and then fine-tuning them on specific tasks.[17]

Finally, even if you're not explicitly aiming for or anticipating agentic planning and strategic awareness in an AI system, these properties could in principle arise in unexpected ways regardless, and/or prove difficult to prevent. For example, optimizing a system to perform some not-intuitively-agential task (for example, predicting strings of text) could, given sufficient cognitive sophistication, result in internal computation that makes and executes plans, in pursuit of objectives, on the basis of broad and informed models of the real world. Indeed, the likelihood of this seems correlated with the strength of the 'usefulness' consideration discussed above: insofar as agentic planning and broad-scope world-modeling are very useful types of cognition, we might expect to see them cropping up a lot in sufficiently optimized systems, whether we want them to or not.

Of these three reasons to expect APS systems—their usefulness, the pressures exerted by available techniques, and the possibility that they arise as byproducts of sophistication—I place the most weight on the first.

5 Alignment

Let's assume that it will become possible and financially feasible to create APS systems before 2070, and that there will be significant incentives to do so. This section argues that it will be difficult to create APS systems that don't seek to gain and maintain power in unintended ways.

5.1 Some definitions

I'll define 'misaligned behavior', in an AI system, as unintended behavior that arises in virtue of problems with an AI system's objectives.[18] A characteristic feature of misaligned behavior is that it is *unintended* but still *competent*. That is, it looks less like an AI system breaking or

[16] This may be especially true of tasks where we lack lots of training data, which could be the majority of useful tasks.

[17] Richard Ngo suggested this point in conversation; see also his discussion of the 'generalization-based approach' in his (2020). The training and fine-tuning used for current large language models may also be suggestive of patterns of this type.

[18] There are ambiguities about what sort of behavior counts as 'intended' by designers (in particular, the relationship between 'intended' and 'foreseen' is unclear), but I'm going to leave the notion vague for now, and assume that behaviors like lying, stealing money, resisting shut-down by appropriate channels, harming humans, and so forth are generally 'unintended'. Similarly, I don't, at present, have a rigorous account of how to attribute unintended behavior to problems with objectives vs. other problems; and I doubt the distinction will always be deep or easily drawn (this doesn't make it useless). But I'll lean on the intuitive notion for now.

failing in its efforts to do what designers want, and more like an AI system trying, and perhaps succeeding, to do something designers *don't* want it to do.[19]

Not all misaligned AI behavior seems relevant to existential risk. Consider, for example, an APS AI system in charge of an electrical grid, whose designers intend it to send electricity to both town A and town B, but whose objectives have problems that cause it, during particular sorts of storms, to only send electricity to town A. This is misaligned behavior, and it may be quite harmful, but it poses no threat to the entire future.

Rather, the type of misaligned AI behavior that I think creates the most existential risk involves misaligned *power-seeking* in particular: that is, active efforts by an AI system to gain and maintain power in ways that designers didn't intend, arising from problems with that system's objectives (I'll call this sort of behavior 'PS-misaligned').[20] AI systems that don't seek to gain or maintain power may cause a lot of harm, but this harm is more easily limited by the power they already have. And such systems, by hypothesis, won't try to maintain that power if/when humans try to stop them. Hence, it's much harder to see why humans would fail to notice, contain, and correct the problem before it reaches an existential scale.[21]

I'll say that a system is 'fully aligned' if it doesn't engage in misaligned behavior on any inputs compatible with the basic physical conditions of our universe (I'll call these 'physics-compatible inputs'),[22] and 'practically aligned' if it doesn't engage in misaligned behavior on any of the inputs it will in fact receive.[23] And I'll say the system is 'fully *PS*-aligned' if it doesn't engage in misaligned *power-seeking* on any physics-compatible inputs, and 'practically *PS*-aligned' if it doesn't engage in misaligned *power-seeking* on any of the inputs it will in fact receive.[24] For our purposes, it is this last property—practical PS-alignment—that matters most.

[19] In this sense, it's less like a nuclear plant melting down, and more like a heat-seeking missile pursuing the wrong target; less like an employee giving a bad presentation, and more like an employee stealing money from the company.

[20] In the electrical grid case, the AI system hasn't been described as trying to gain power (for example, by trying to hack into more computing resources to better calculate how to get electricity to town A) or to maintain the power it already has (for example, by resisting human efforts to remove its influence over the grid). And in this sense, I think, it's much less dangerous.

[21] That said, note that not all misaligned power-seeking, even in APS systems, is particularly harmful, or intuitively worrying from an existential risk perspective. In a comment on a draft of Carlsmith (2022), Ben Garfinkel suggested the example of a robo-cop pinning down the wrong person, but remaining amenable to human instruction otherwise. And in general, it seems most dangerous if an APS system is *using* its advanced capabilities in pursuing power; e.g. if I'm a great hacker, but a poor writer, and I'm trying to get power via my journalism, I'm less threatening.

[22] Thus, it is physics-compatible for a randomly chosen bridge in Indiana to get hit, on a randomly chosen millisecond in May 2050, by a nuclear bomb, a 10 km-diameter asteroid, and a lightning bolt all at once, but not for the laws of physics to change, or for us all to be instantaneously transported to a galaxy far away. Obviously the scope here is very broad, but note that misaligned behavior is a different standard than 'bad' or even 'catastrophic' behavior. It will always be possible to set up physics-compatible inputs where a system makes a mistake, or gets deceived, or acts in a way that results in catastrophic outcomes. To be misaligned, though, this behavior needs to arise from problems with the system's objectives in particular. Thus, for example, if Bob is a paper clip maximizer, and he builds Fred, who is also a paper clip maximizer, Fred will (on my definition) be fully-aligned with Bob as long as Fred keeps trying to maximize paper clips on all physics-compatible inputs (even though some of those inputs are such that trying to maximize paper clips actually minimizes them, kills Bob, etc.). Thanks to Eliezer Yudkowsky, Rohin Shah, and Evan Hubinger for comments on the relevant scope here (which isn't to say they endorse my choice of definition).

[23] By 'inputs', I mean information the system receives via the channels intended by its designers (an input to GPT-3, for example, would be a text prompt). I am not including processes that intervene in some other way on the internal state of the system—for example, by directly changing the weights in a neural network (analogy: a soldier's loyalty need not withstand arbitrary types of brain surgery).

[24] That is, and importantly, for a system to be practically PS-aligned, it doesn't need to be the case that it would never, in any circumstances (or with any level of capability), engage in problematic power-seeking. This is in contrast with some other strands of the literature. See e.g. Yudkowsky (n.d.c) on the 'omnipotence test for AI

5.2 Power-seeking

Why might we think that it will be hard to prevent misaligned power-seeking? A key hypothesis, some variant of which underlies much of the discourse about existential risk from AI, is that there is a close connection (in sufficiently advanced agents) between misaligned behavior in general and misaligned power-seeking in particular.[25] I'll formulate this hypothesis as follows:

> *Instrumental convergence*: If an APS AI system is less-than-fully aligned, and some of its misaligned behavior involves strategically aware agentic planning in pursuit of problematic objectives, then in general and by default, we should expect it to be less-than-fully PS-aligned, too.[26]

Why believe in *instrumental convergence*? The basic reason is that power, almost by definition, is extremely useful to accomplishing objectives. So to the extent that an agent is engaging in unintended behavior in pursuit of problematic objectives, it will generally have incentives, other things equal, to gain and maintain forms of power in the process—incentives that strategically aware agentic planning puts it in a position to recognize and respond to.[27]

What sorts of power might a system seek? Bostrom (2014) identifies a number of 'convergent instrumental goals', each of which promotes an agent's power to achieve its objectives.[28] These include:

- self-preservation (since an agent's ongoing existence tends to promote the realization of its objectives);

safety': 'The Omni Test is that an advanced AI should be expected to remain aligned, or not lead to catastrophic outcomes, or fail safely, even if it suddenly knows all facts and can directly ordain any possible outcome as an immediate choice. The policy proposal is that, among agents meant to act in the rich real world, any predicted behavior where the agent might act destructively if given unlimited power (rather than e.g. pausing for a safe user query) should be treated as a bug.' (See Yudkowsky (n.d.a.; n.d.e) in 'AI safety mindset', 'Querying the AGI user'.) Talk of an 'objective' such that the 'optimal policy' on that objective leads to good outcomes is also reminiscent of something like the Omni Test. See e.g. Hubinger's (2020) definition of 'intent alignment': 'An agent is intent aligned if the optimal policy for its behavioral objective is aligned with humans.' Also see Christiano (2018) and Hubinger et al. (2019: 35).

[25] See e.g. Bostrom (2014: 127–40) and Russell (2019: 132–45).

[26] Note that instrumental convergence is not a conceptual claim, but rather an empirical claim that purports to apply to a wide variety of APS systems. In principle, for example, we can imagine APS systems that plan in pursuit of problematic objectives on some inputs, but which are nevertheless fully PS-aligned (or very close to it). Consider, for example, an APS version of the electrical grid AI system above, which plans strategically in pursuit of directing electricity only to town A, but which just doesn't consider plans that involve seeking power. That said, the in-principle possibility of strategic, agentic misalignment without PS-misalignment is important, though, since it might be realized in practice. Perhaps, for example, the type of training we should expect by default will reinforce cognitive habits in APS systems that steer away from searching over/evaluating plans that involve misaligned power-seeking, even if other types of misaligned behavior persist.

[27] One way to bring this out is to conceptualize power in terms of the number of options an agent has available to it. Thus, if a policy seeks to promote some outcomes over others, then other things equal, a larger number of options makes it more likely that a more preferred outcome is accessible. Indeed, talking about 'options-seeking', instead of 'power-seeking', might have less misleading connotations.

[28] Bostrom is following Omohundro (2008). Here I am thinking of the pursuit of any of these instrumental goals as a part of 'power-seeking' in the relevant sense.

- preventing changes to its objectives (since agent's pursuit of those objectives in particular tends to promote them);
- improving its cognitive capability (since such capability tends to increase an agent's success in pursuing its objectives);
- technological development (since control over more powerful technology tends to be useful);
- resource acquisition (since more resources tend to be useful, too).

We've already seen examples of rudimentary AI systems 'discovering' the usefulness of resource acquisition, for example. When OpenAI trained two teams of AIs to play hide and seek in a simulated environment that included blocks and ramps that the AIs could move around and fix in place, the AIs learned strategies that depended crucially on acquiring control of the blocks and ramps in question—despite the fact that they were not given any direct incentives to interact with those objects (the hiders were simply rewarded for avoiding being seen by the seekers; the seekers, for seeing the hiders).[29]

Of course, this is a very simple, simulated environment, and the level of agentic planning it makes sense to ascribe to these rudimentary AIs isn't clear.[30] But the basic dynamic here applies in the real world, as well, and it applies to more advanced systems. Things like money, compute, energy, and social influence will often help agents achieve their objectives. We should expect a sufficiently sophisticated, strategically aware AI agent to register this fact, and make its decisions accordingly. Thus a sophisticated AI system might try to hack into financial databases; take control of compute; manipulate human opinion; and so on.[31]

One objection to instrumental convergence is that many humans don't seem particularly 'power-seeking', despite their agentic planning and strategic awareness.[32] And it's true that many humans do not seek various types of power *in their current circumstances*—in which (for example) their capabilities are roughly similar to those of their peers, they are subject

[29] Thus, the hiders learned to move and lock blocks to prevent the seekers from entering the room where the hiders were hiding; the seekers, in response, learned to move a ramp to give them access anyway; adjusting for this, the hiders learned to take control of that ramp before the seekers could get to it, and to lock it in the room as well. In another environment, the seekers learned to 'surf' on boxes, and the hiders, to prevent this, learned to lock *all* boxes and ramps before the seekers could get to them. For more detail see Baker et al. (2020).

[30] And importantly, the trainers weren't *trying* to disincentivize resource-seeking behavior; quite the contrary, the set-up seems (I haven't investigated the history of the experiments) to have been designed to test whether 'emergent tool-use' would occur.

[31] Other concrete examples of unintended power-seeking might include AI systems trying to: break out of a contained environment; make backup copies of themselves; gain unauthorized capabilities, sources of information, or channels of influence; mislead/lie to humans about their goals; resist or manipulate attempts to retrain them or shut them off; create/train new AI systems themselves; coordinate illicitly with other AI systems; impersonate humans; cause humans to do things for them; increase human reliance on them; weaken various human institutions and response capacities; take control of physical infrastructure like factories or scientific laboratories; cause certain types of technology and infrastructure to be developed; or directly harm/overpower humans.

[32] Humans care a lot about survival, and about certain resources (food, shelter, etc.), but beyond that, we associate many forms of power-seeking with a certain kind of greed, ambition, or voraciousness, and with intuitively 'resource-hungry' goals, like 'maximize X across all of space and time'. *Some* strategically aware human planners are like this, we might think, but not all of them: so strategically aware, agentic planning isn't, itself, the problem. For more in this vein, see e.g. Cegłowski's (2016) 'Argument from My Roommate'; and also Pinker (2018: 297): 'There is no law of complex systems that says that intelligent agents must turn into ruthless conquistadors. Indeed, we know of one highly advanced form of intelligence that evolved without this defect. They're called women.' Thanks to Rohin Shah for discussion of the humans example.

to various social and legal incentives, they are hemmed in by significant physical and temporal constraints, and they recognize certain intrinsically important ethical constraints. But almost all humans will seek to gain and maintain various types of power in *some* circumstances, especially when they can do so at little cost. Thus, for most humans, it doesn't make sense to try to start a billion-dollar company—the expected returns on such effort are too low. But most humans will walk across the street to pick up a billion-dollar check. More generally, the power-seeking behavior humans display, when getting power is easy, seems to me quite compatible with the instrumental convergence thesis. And unchecked by ethics, constraints, and incentives (indeed, even *when* checked by these things) human power-seeking seems to me plenty dangerous, too.[33]

A second objection (in possible tension with the first) is: *humans* (or, some humans) may be power-seeking, but this is a product of a specific evolutionary history (namely, one in which things like survival, resource-acquisition, and social dominance were directly selected for), which AI systems will not share.[34] Some versions of this objection simply neglect to address the instrumental convergence argument above[35] (and note, regardless, that some proposed ways of training AI systems resemble evolution in various respects).[36] But we can see stronger versions as suggesting that maybe it just isn't that hard to train APS systems not to seek power in unintended ways, across a large enough range of inputs, if you're actively *trying* to do so (evolution wasn't). And I do think this is possible—indeed, it's one of my main sources of hope. But I also think that there are barriers to overcome, which I discuss in the next section.

There's more to say about the instrumental convergence thesis.[37] The formulation I've offered here might warrant refinement—and indeed, this part of the argument is one of my top candidates for ways that the abstractions employed might mislead.[38] Still, I expect the basic gist to point at something real—the main question is how often it arises, and how difficult it is to avoid.

[33] That said, the absence of various forms of overt power-seeking in humans may point to ways we could try to maintain control over less-than-fully PS-aligned APS systems.

[34] See e.g. Zador and LeCun (2019) (and follow-up debate Pace (2019)), and Pinker (2018: ch. 19). One can also imagine non-evolutionary versions of this—e.g. ones that attribute human power-seeking tendencies to our culture, our economic system, and so forth. Indeed, Cegłowski (2016) can be read as suggesting something like this in the context of a particular demographic: after listing Bostrom's convergent instrumental goals, he writes: 'If you look at AI believers in Silicon Valley, this is the quasi-sociopathic checklist they themselves seem to be working from.'

[35] That is, the argument isn't 'humans seek power, therefore AIs will too'; it's 'power is useful for pursuing objectives, so AIs pursuing problematic objectives will have incentives to seek power, by default.'

[36] Large, multi-agent reinforcement learning environments might be one example. And the 'league' used to train AlphaStar (see Vinyals et al. 2019: 350–54) seems reminiscent of evolution in various ways.

[37] See Carlsmith (2020: sec. 4.2) for more details.

[38] In particular, this part of the argument requires that the agentic planning and strategic awareness at stake be robust enough to license predictions of the form: 'if (i) a system would be planning in pursuit of problematic objectives in circumstance C, (ii) power-seeking in C would promote its objectives, and (iii) the models it uses in planning put it in a position to recognize this, then we should expect power-seeking in C by default.' I've tried to build something like the validity of such predictions into my definitions of agentic planning and strategic awareness; but perhaps for sufficiently weak/loose versions of those concepts, such predictions are not warranted; and it seems possible to conflate weaker vs. stronger concepts at different points in one's reasoning, and/or to apply such concepts in contexts where they confuse rather than clarify. Thanks to Rohin Shah for emphasizing possible objections in this vein, and for discussion.

5.3 The challenge of practical PS-alignment

Let's grant that less-than-fully aligned APS systems will have at least some tendency toward misaligned, power-seeking behavior, by default. The challenge, then, is to prevent such behavior in practice—whether through alignment (including full alignment), or other means. How difficult will this be?

We can group possible interventions here into three types. The first attempts to control a system's *objectives*. The second attempts to control its *capabilities*. The third attempts to control its *circumstances*. Each type has problems.

5.3.1 Controlling objectives

Much current work on technical alignment focuses on shaping an AI system's objectives to prevent misaligned power-seeking.[39] But this project must confront at least two major obstacles.

5.3.1.1 Problems with proxies

Many ways of attempting to control an AI's objectives share a common challenge: namely, that giving an AI system a 'proxy objective'—that is, an objective that reflects properties correlated with, but separable from, intended behavior—can result in behavior that weakens or breaks that correlation, especially as the power of the AI's optimization for the proxy increases.

We're familiar with this problem in human contexts. To give just one example: suppose I try to lower the cobra population by paying the people in my town a bounty for each dead cobra. Initially they might kill lots of existing cobras. But as the cobra population decreases, they might also start to *breed* cobras, in order to kill them and turn them in.[40] Optimizing for 'turn in dead cobras' breaks that proxy objective's correlation with 'reduce the cobra population', which is what I actually want the townspeople to do.

The same phenomenon can arise in training AIs. Suppose, for example, that I'm trying to get an AI to engage in some behavior I want, by rating its behavior using some type of feedback. The correlation between the behavior I would rate highly and the behavior I *actually* want will be strong so long as I can monitor and understand the AI's behavior. But if the AI is too sophisticated for me to understand everything it's doing, and/or if it can deceive me about its action, the correlation weakens: the AI may be able to cause me to give high ratings to behavior I wouldn't (in my current state) endorse if I understood it better—for example, by hiding information about that behavior, or by manipulating my preferences.

[39] Note that the mechanisms we have available to shaping an AI system's objectives change over time. I emphasize this because sometimes the challenge of AI alignment is framed as one of shaping an AI's objectives *in a particular way*—for example, via hand-written code, or via some sort of reward signal, or via English-language sentences that will be interpreted in literalistic and uncharitable terms. And this can make it seem like the challenge is centrally one of e.g. coding, measuring, or articulating explicitly everything we value, or getting AI systems to interpret instructions in commonsensical ways. These challenges may be relevant in some cases, but the core problem is not method-specific.

[40] Inspired by a (possibly fictitious) anecdote described in a Wikipedia entry. Other examples: paying railroad builders by the mile of track that they lay incentivizes them to lay unnecessary track ('Perverse incentive': 'Returns for effort'); if teachers take 'cause my students to get high scores on standardized tests' as their objective, they're incentivized to 'teach to the test'—an incentive that can work to the detriment of student education more broadly; and so on. See Wikipedia ('Perverse incentive') for more examples. See Manheim and Garrabrant (2019) for an abstract categorization of dynamics of this kind.

We already see this sort of problem in existing AI systems. Thus, for example, if we train an AI system to complete a boat race by rewarding it for hitting green blocks along the path to the finish line, it learns to drive the boat in circles in order to hit the same green blocks over and over again (see Clark and Amodei: 2016).[41] Examples like these may seem easy to fix, but they illustrate the more general problem: systems optimizing for proxies often behave in unintended ways. Indeed, this tendency is closely connected to a core property that makes advanced AI useful: namely, the ability to find novel solutions and strategies that humans wouldn't think of.[42]

How can we address this problem? Human feedback seems likely to play a key role.[43] And it may, ultimately, be enough. But notably, we need ways of drawing on this feedback that don't require unrealistic amounts of human supervision and human-generated data;[44] we need to ensure that such feedback captures our preferences about behavior that we can't directly understand and/or whose consequences we haven't yet seen;[45] we need ways of eliminating incentives to manipulate or mislead the human feedback mechanisms in question; and we need such methods to scale competitively as frontier AI capabilities increase.

Would it help if our AI systems could understand fuzzy human concepts like 'helpfulness', 'obedience', 'what humans would want', and so forth? I expect it would, in various ways (though as I discuss below, this also opens up new opportunities for deception/manipulation). But note that the key issue isn't getting our AI systems to *understand* what objectives we want them to pursue—indeed, such understanding is plausibly on the critical path to increasing their capability, regardless of their alignment.[46] Rather, the key issue is causing them to *pursue* those objectives for their own sake.[47]

5.3.1.2 Problems with search

Many techniques shape an AI's objectives using proxies. But some techniques—namely, those that involve *searching* over AI systems and selecting those that perform well on some evaluation criteria, without controlling the systems' objectives directly—face an additional problem.

The problem is that, *even if* the search criteria fully capture the behavior we want, the resulting systems may not end up intrinsically motivated by the criteria in question. Instead, they may end up with other objectives, pursuit of which correlated with good performance

[41] See Krakovna et al. (2020) for a much longer list of examples in this vein.

[42] When you don't know how an AI will achieve its objective, and that objective doesn't capture everything that you really want, then even for comparatively weak systems and simple tasks, it's hard to anticipate how the system's way of achieving the objective will break its correlation with what you really want. And as the AI's capacity to generate solutions we can't anticipate grows, the problem becomes more and more challenging.

[43] This paragraph draws heavily on discussion in Ngo (2020).

[44] See e.g. Christiano et al (2017).

[45] See Christiano, Shlegeris, and Amodei (2018) for discussion. Iterative amplification and distillation (Christiano et al. 2018), debate (Irving, Christiano, and Amodei 2018), and recursive reward modeling (Leike et al. 2018) can all be seen as efforts in this vein.

[46] This is a point from Ord (2020). For example, if we train some set of sophisticated agents to get bananas, in a complex environment that requires understanding and modeling humans, they may end up capable of understanding quite accurately (even more accurately than us) what we have in mind when we talk about 'aligned behavior', and of behaving accordingly (for example, when we give them bananas for doing so). But their intrinsic objectives could still be focused centrally on bananas (or something else), and our abilities to control those objectives directly might remain quite limited.

[47] Though if they understand those objectives, but don't share them, we might also be able to incentivize them to pursue such objectives for instrumental reasons.

during the selection process, but which lead to unintended behavior when the system is exposed to other inputs.

Some think of human evolution as an example.[48] Someone interested in creating agents who pass on their genes could run a process similar to evolution, which searches over different agents, and selects for ones who pass on their genes (for example, by allowing ones who don't to die out). But this doesn't mean the resulting agents will be intrinsically motivated to pass on their genes. Humans, for example, are motivated by objectives that were *correlated* with passing on genes (for example, avoiding bodily harm, having sex, securing social status, etc.), but which they'll pursue in a manner that breaks such correlations, given the opportunity (for example, by using birth control, or remaining childless to further their careers).

Rudimentary, evolved AI systems display analogous tendencies. Thus, when Ackley and Littman ran an evolutionary selection process in an environment with trees that allowed agents to hide from predators, the agents developed such a strong attraction to trees that (after reproductive age) they would starve to death in order to avoid leaving tree areas.[49] And we see early evidence of similar dynamics in systems trained via gradient descent.[50] Thus, for example, when an agent is given reward for visiting colored spheres in a certain order, and trained in an environment where an 'expert' traces the correct path, it learns to follow the expert rather than to trace the correct path—thus accumulating large amounts of negative reward when paired with an 'anti-expert' who traces the path incorrectly.

The problem, in these various cases, isn't that the evaluation criteria (e.g. 'pass on genes', 'visit the spheres in X order') fail to fully capture and operationalize what we want. The problem is that selecting agents by reference to these evaluation criteria doesn't afford the designers enough control over the objectives of the resulting agents.

It's an open empirical question how often problems like this will arise. But there are reasons to worry. For one thing, proxy goals correlated with evaluation criteria may be simpler, and therefore easier to learn, than the evaluation criteria.[51] What's more, if the 'actual' objective function provides slower feedback, agents that pursue faster-feedback proxies may have advantages.[52] Finally, to the extent that many objectives would *instrumentally* incentivize good behavior in training (for example, because many objectives, when coupled with strategic awareness, incentivize gaining power in the world, and doing well in training leads to deployment/greater power in the world), but few involve *intrinsic* motivation to engage in such behavior, we might think it more likely that selecting for good behavior leads to agents who behave well for instrumental reasons.

[48] See e.g. Hubinger et al. (2019: 6).

[49] See Christian (2020).

[50] See Shah et al. (2022) and Langosco et al. (2023) on examples of 'goal misgeneralization'.

[51] In the context of evolution, for example, it seems much harder to evolve an agent whose mind represents a concept like 'passing on my genes', and then takes doing this as its explicit goal—humans, after all, didn't even have the concept of 'genes' until very recently—than to evolve an agent whose objectives reflect the relevance of things like bodily damage, sex, power, knowledge, etc. to whether its genes get passed on (though starting with cognitively sophisticated agents might help in this respect). This is a point I believe I heard first from Evan Hubinger.

[52] For example: in the game Montezuma's Revenge, it helps to give an agent a direct incentive analogous to 'curiosity' (e.g. it receives reward for finding sensory data it can't predict very well), because the game's 'true' objective (e.g. exiting a level by finding keys that require a large number of correct sequential steps to reach) is too difficult to train on. See Burda et al. (2018) and discussion in Christian (2020).

Overall, ensuring robust practical PS-alignment seems harder if available techniques search over systems that meet some external evaluation criteria, with little direct control over their objectives. And much of contemporary machine learning fits this bill.

5.3.1.3 Myopia

Some broad types of objectives seem to incentivize power-seeking on fewer physics-compatible inputs than others. Perhaps, then, we can aim at those, even if we lack more fine-grained control.

Short-term (or, 'myopic') objectives seem especially interesting here.[53] Paradigmatically dangerous AI systems plan in pursuit of long-term objectives. Longer time horizons allow more time to gain and use forms of power that aren't readily available, and more easily justify temporarily costly action (for example, trying to appear aligned, in order to get deployed) for the sake of longer-term gains. Since myopic agents are on a much tighter schedule, they have weaker incentives to attempt forms of power-seeking (deception, resource acquisition, etc.) that only pay off in the long run.[54]

Myopia might help, but I see at least two problems with relying on it. First, there will plausibly be demand for non-myopic agents. Human individuals and institutions often have fairly (though not arbitrarily) long-term objectives that require long-term planning—running factories and companies, pursuing electoral wins, and so on. As I've already discussed, there will be powerful incentives to automate the pursuit of these objectives. Non-myopic systems will have an advantage, in this regard, over myopic ones.

Second, the 'search' techniques discussed in the previous section—techniques that don't allow you to control an agent's objectives directly—may make ensuring myopia challenging. And various types of long-term training processes—for example, reinforcement learning on tasks that involve many sequential steps—seem likely to result in non-myopia by default.[55]

5.3.2 Controlling capabilities

AI alignment research often focuses on controlling a system's objectives. But controlling its capabilities can help with practical PS-alignment too. The less capable a system, the more easily its behavior (including its tendencies toward misaligned power-seeking) can be anticipated and corrected. Less capable systems will also have a harder time getting and keeping power, and a harder time making use of it. So they will have stronger incentives to cooperate with humans, rather than (say) deceive or overpower them.

Preventing agentic planning and strategic awareness in the first place would be one example of 'controlling capabilities', but there are other options, too. I'll consider two.

[53] We can also imagine other strategies for controlling shrinking the set of inputs that prompt PS-misaligned behavior. For example, we might aim for objectives that penalize 'high-impact' action (see e.g. the discussion of 'impact penalties' in Krakovna et al. (2020)), or that prohibit lying in particular, or that give intrinsic weight to various legal and ethical constraints. But these face the same challenges involving proxies and search discussed in the last two sections.

[54] Of course, even short spans of time can be enough to do a lot of harm, especially for extremely capable systems. And the time spans 'short enough to be safe' can alter if what one can do in a given span of time changes. Thanks to Rohin Shah, Paul Christiano, and Carl Shulman for discussion.

[55] That said, myopia is a fairly coarse-grained property for an objective to possess, and may be easier to cause or check for than others.

The first is *specialization*: we might try to build APS systems whose competence is as narrow and specialized as possible. After all, an APS system skilled at a specific kind of scientific research, and not at (say) hacking and social persuasion, seems much less dangerous than an APS system that can engage in these other tasks as well. And specialization often has various benefits (hence its importance in human organizations and economies).[56] That said, generality has benefits, too—which is why human workers with quite general skill-sets (CEOs, for example) are prized in many domains.[57] And just as available machine learning techniques may push the field toward agentic planning and strategic awareness, so too they may push it toward generality.[58] Plus, even very specialized APS systems can be dangerous.[59]

The second strategy is *preventing problematic improvements in capability*. New capabilities can put a system in a position to gain and maintain power in ways it couldn't before—and hence, make new incentives action-relevant. Practical PS-alignment may therefore require controlling the extent to which the inputs a system receives result in improved capabilities. This seems easier if the variables in the system that determine how it responds to inputs (for example, the weights in a neural network) stay fixed. But we may also want systems that mix task-performance and learning together, that 'remember' previous events, and so forth; and predicting and controlling the capabilities such systems will develop could be difficult (especially if we don't understand well how they work—more below in subsection 5.4).

Strategies that rely on limiting a system's capabilities also face a more general problem: namely, that there are likely to be strong incentives to scale up the capabilities of frontier systems.[60] PS-alignment strategies that can't scale accordingly (and competitively) therefore risk obsolescence as state-of-the-art capabilities advance. A key question for any such strategy, then, is whether it can translate, given success at some level of capability, into a different strategy that scales better.[61]

[56] For example, they can be optimized more heavily for specific functions (to borrow an (unpublished) example from Ben Garfinkel, there is a reason that the flashlight, camera, speakers, etc. on an iPhone are inferior to the best flashlights, cameras, etc.). And note that we will have much greater abilities to optimize AI systems for particular tasks than we do with humans.

[57] In particular, general systems plausibly respond better to changing environments and demands; and if a task requires multiple competencies, specialized systems can be harder to coordinate (e.g. it's helpful to have a single personal assistant, rather than one for email, one for scheduling, one for travel planning, one for research, etc.).

[58] GPT-3, for example, is trained to a fairly general level of capability via predicting text, and later fine-tuned on specific tasks like coding. Such an approach might be necessary for tasks where data is hard to come by or learn from directly (e.g. 'be an effective CEO'). More broadly, as Bostrom (2014) notes, the most efficient route to widespread automation may be the creation of general purpose agents that can learn a wide variety of new tasks very efficiently (though those agents could also end up quite specialized later).

[59] A system highly skilled at hacking into new computers and copying itself, for example, can spread far and wide; a system skilled in science can design a novel virus; a system with control over automated weapons can use them; a system skilled at social manipulation can turn an election; and so forth.

[60] Though note that PS-alignment problems with more capable systems could complicate these dynamics. More discussion at the beginning of section 6 below.

[61] For example, we might try to achieve practical PS-alignment with some fairly advanced systems (including, perhaps, quite specialized ones—or, indeed, non-APS ones), and then use them to create new and superior PS-alignment strategies (indeed, as AI development itself becomes increasingly automated, automating alignment research will plausibly be necessary regardless). Note, though, that plans of the form 'create some practically PS-aligned systems, and ask them what the plan should be' might just not work. For example, the new systems might not have adequate plans either. One might therefore need to create even more capable systems, which *also* might not have adequate plans, and so forth, until one pushes up against (or perhaps, past) the limits of one's capacity to ensure practical PS-alignment.

5.3.3 Controlling circumstances

Alongside (or instead of) controlling a system's 'internal' properties—its objectives or capabilities—we might also try to control its *external circumstances*.[62] If we want to prevent a system from engaging in hacking, for example, we might try to control its options (no internet access for the system); its incentives (if the system is caught hacking it will be disabled, thus reducing its ability to pursue its objectives); or both.

The success of this strategy will depend a lot on the system's capabilities. If it is sufficiently good at hacking, our efforts to prevent it from accessing the company bank account might fail. If the system is sufficiently good at avoiding detection, our threats to disable it if we catch it hacking will provide little incentive for it to refrain. So in order to contain problematic power-seeking by controlling options or incentives, we need monitoring and enforcement mechanisms that scale with the capabilities of frontier APS systems. But this may prove very difficult; as the capabilities of frontier systems increase, their capacity to evade and disable our monitoring and enforcement mechanisms will increase as well.[63]

5.4 Unusual difficulties

So all three strategies for ensuring PS-aligned behavior we've considered—controlling objectives, capabilities, and circumstances—face serious difficulties. Still, one might suggest, we have compelling inductive grounds for thinking that these difficulties can be overcome. After all, plenty of technologies encounter serious safety problems in the course of their development, but we often reach adequate safety standards in the end. Planes don't generally crash; bridges stay up; email servers are tolerably secure; and so on.

But APS AI systems pose safety challenges that (most) other technologies do not. Here are a few that seem salient.

The first is *opacity*. Ensuring safety and reliability requires understanding a system well enough to predict its behavior. But the thinking and strategic decision-making of an agent whose cognitive capacities significantly exceed our own may be quite alien and opaque to us. Such a system might consider options humans never would; understand physical and social dynamics that humans do not; reason in ways we cannot understand, and so on.[64]

This issue seems especially salient in the current, machine-learning-dominated AI paradigm, in which our ability to create an AI system that can perform some task (e.g. predicting text) often far exceeds our ability to understand *how* the system does what it does. We set

[62] Admittedly, the line between capabilities and circumstances is somewhat blurry, but I mean for the former to lean more heavily on internal features of the system, and the latter, on external features.

[63] Also, ensuring practical PS-alignment seems harder insofar as a system's deployment circumstances are dissimilar from the ones on which humans have observed and verified PS-aligned behavior (for example, during training, or pre-deployment testing). Indeed, ideally, one would want the deployment inputs to come from the same distribution as the training inputs. But in practice, and especially in strategically aware systems, ensuring a close-to-identical distribution seems very difficult (if not impossible). This is partly because the world changes (and indeed, the actions of the APS systems can themselves change it. See Christian (2020) and Krueger, Maharaj, and Leike (2020) for some discussion of this possibility.) But also, to the extent that the distinction between training and deployment reflects some real difference in the agent's level of influence on the world, this difference is itself a change in distribution—one that a sufficiently sophisticated agent might recognize.

[64] See Yudkowsky (n.d.d) on 'strong cognitive uncontainability'.

various key high-level variables (the system's architecture, the number of parameters, the training process, the evaluation criteria), but the system that results is still, in many (though not all) respects, a black box. We must rely on further experiments to try to get some handle on what it knows, what it can do, and how it is liable to behave.[65] This marks an important contrast with technologies like planes and bridges, where we achieve safety partly through understanding of the basic physical principles that govern their behavior.[66]

The second challenge comes from *adversarial dynamics*, especially in the context of efforts to detect safety problems. Suppose an AI system has an objective that it can better achieve by passing a training or evaluation process. It may then manipulate that process, or deceive the people conducting it.[67] And if it is sufficiently capable at this, the appearance of safety and reliability on various tests may tell us little about how the system will behave in other circumstances. Few if any existing technologies exhibit this dynamic. Planes, rockets, and nuclear plants may be dangerous and complicated, but they never *try* to appear safer than they are, or to manipulate our ability to understand and evaluate them.

A final challenge involves the *stakes of error*. If an engineered virus escapes from a lab, it can spread rapidly and become increasingly difficult to contain. And this seems like a much better analogy for a misaligned, power-seeking AI system than a plane crashing or a bridge falling down. Because the stakes of error are so high, there is much less room for trial and error.[68] Indeed, if you're trying to store an engineered virus that has a significant chance of killing most of the global population if it gets released, you need safety standards *much* higher than those we use, even now (after generations of trial and error), for bridges or planes—much higher, indeed, than we use for approximately anything (this is one key reason to never, ever create such a virus). And humanity's track record in the highest stakes contexts—biosafety labs that handle the most dangerous viruses, nuclear power plants, nuclear weapons facilities—seems far from comforting.[69]

5.5 Overall difficulty

Overall, then, ensuring practical PS-alignment seems like it could well prove challenging. And while there are tools that might help—myopia, restricting capabilities, and so on—there are significant problems with each of these tools, along with more general reasons to think the problem uniquely difficult relative to technological safety problems we've faced in the past.

[65] Of course, our understanding of how machine learning systems work will likely improve over time, and research in this area ('interpretability') is ongoing (see e.g. Olah et al. (2020) and Goh et. al. (2021)). But interpretability is no bottleneck to training bigger models on even more complex tasks—or, plausibly, to the commercial viability of those models. And much of the AI field is devoted to pushing forward with developing whatever capabilities we can, interpretable or no.

[66] That said, understanding comes in many varieties and degrees; and it's an empirical question what mix of experiment/search vs. first-principles understanding/design has actually been involved in ensuring the safety of different technologies (for example, in biology, or before advanced scientific understanding).

[67] Bostrom (2014) calls this a 'treacherous turn'. See also Cotra (2021) on the 'training game'.

[68] And whatever their present safety, most current technologies involved many errors (plane crashes, rocket explosions, etc.) along the way.

[69] See Ord (2020: 130) for some discussion of biological accidents in particular.

6 Deployment

Let's suppose, then, that ensuring practical PS-alignment is difficult. Should we expect to see practically PS-misaligned APS systems actually deployed?

One might think: no. After all, if we couldn't figure out how to build planes that don't crash, we wouldn't expect to see people dying in plane crashes all the time. Rather, we'd expect to see people not flying. What's more, plenty of mundane commercial incentives favor safety (safety failures can result in significant social/regulatory backlash and economic cost), and the incentives to prevent harmful, large-scale forms of misaligned power-seeking seem especially clear (since sufficiently severe failures can result in the direct disempowerment of the relevant decision-makers, their loved ones, and so on).

Faced with such incentives, why would anyone deploy a strategically aware AI agent that will seek power in unintended ways? A simple reason is: the decision-maker might just act stupidly, and contrary to the evidence and incentives a rational decision would reflect. But I think there are more specific reasons for concern. Here I'll focus on four.[70]

The first is the familiar phenomenon of *externalities*. It might be individually rational for a less-than-fully altruistic actor to deploy a possibly PS-misaligned system, even if that's very bad in expectation for society, if society's interests (let alone: the interests of all future generations) are inadequately reflected in the actor's incentives.[71] Indeed, even if the actor recognizes the risks to society's interests, and gives them some weight, he/she might not give them *enough* weight to outweigh the prospect of personal profit, power, or prestige.[72]

The second and related reason is *race dynamics*. The time and effort that an AI developer devotes to ensuring practical PS-alignment trades off against the speed with which she can scale up its capabilities. And there are significant rewards to deploying such a system before others do.[73] In order to beat her competitors, therefore, a developer might choose speed over safety. And this might then incentivize other developers to take on increased risk as well, creating further incentives for the initial developer to move speedily and riskily. The result might be an ongoing feedback loop of increasing pressure on all parties—altruistic and not-so-altruistic—to ratchet up their risk tolerance or drop out of the race.

The third reason is that there will be *many relevant actors*. That is, once *some* actors can create APS systems, then over time (and absent active efforts to the contrary) a larger and larger number of actors around the world will likely become able to do so as well. And even if some actors are sufficiently conscientious, intelligent, and responsible to avoid problematic deployment (this alone is far from guaranteed), others plausibly won't be. For example, they might be overconfident in the degree of PS-alignment of the systems they've created; or overly dismissive of the amount of harm that misaligned systems would cause; or insufficiently incentivized to register risks to larger society; or insufficiently constrained by legal

[70] An additional risk here is the possibility of 'unintentional deployment'. Suppose, for example, that a power-seeking system meant to be contained in some training environment, or limited in its means of influencing the outside world, manages to break out of that environment, or to obtain other forms of influence. It may then succeed in gaining various forms of real-world power, even though it was never intentionally deployed.

[71] Here we might think of analogies with climate change.

[72] Of course, an actor who deploys a possibly misaligned AI risks harm to *themselves*, as well. But especially if the risk of misalignment-and-resulting-catastrophe is sufficiently small, then it may be in the actor's narrow self-interest to deploy the AI, although the expected disvalue to wider society is very large.

[73] See Askell, Brundage, and Hadfield (2019) for discussion of first-mover examples. Bostrom's notion of a 'decisive strategic advantage' (2014, Chapter 5) is an extreme example.

and regulatory mechanisms, liability regimes, and so on that apply elsewhere. Of course, we can look for ways of setting up coordination and incentive mechanisms that would apply to *all* the relevant actors, around the world (including e.g. China, Russia, North Korea, etc.)—but efforts in this vein seem challenging.[74]

The fourth reason is that practically PS-misaligned APS systems can still be *extremely useful*, at least initially. Obviously, if an APS system is blatantly failing to behave in the way that its designers intend, including during testing and evaluation, then it's much less likely to get deployed (compare with e.g. a house-cleaning robot that routinely kills a user's pets). But practically PS-misaligned APS systems need not behave like this. For one thing, as discussed above, they might actively deceive designers about their degree of alignment.[75] But even absent deception/manipulation of this kind, it's hard to predict/test the AI's behavior on the full range of post-deployment inputs—especially in a rapidly changing world, in the absence of deep understanding of how the system works, and if the AI system might gain new knowledge and capabilities post-deployment. Indeed, I think that one of the central reasons we should expect to see practically PS-misaligned AI systems getting used/deployed is precisely that they will *demonstrate* a high degree of usefulness during training/testing—and consequently, it will be hard to resist deploying them.

Here's an analogy. Suppose that scientists create a new, genetically engineered species of chimpanzee, whose cognitive capabilities significantly exceed those of humans. Initially, scientists confine these chimps in a laboratory environment, and incentivize them to perform various low-stakes intellectual tasks using rewards like food and entertainment. And suppose, further, that these chimpanzees are clearly capable of generating things like vaccine designs, prototypes for new clean energy technology, cures for cancer, highly effective military/political/business strategies, and so forth—and that they will in fact do this, if you set up their incentives right (even though they don't intrinsically value being helpful to humans, and so are disposed, in some circumstances, to seize power for themselves—for example, if they can get more food and entertainment by doing so).

In such a context, it would become increasingly difficult for various actors around the world to resist drawing on the intellectual capabilities of the chimps in a manner that gives

[74] Indeed, to the extent that resources invested in ensuring practical PS-alignment trade off against resources invested in increasing the capabilities of the systems one builds, over time we might expect to see actors who invest less in alignment, and who take more risks, to scale up the capabilities of their systems faster. This could result in the competitive dynamics discussed above (e.g. other actors cut back on safety efforts to keep up, and/or deploy systems that wouldn't meet their own safety standards, but which are safer than the ones they expect competitors to deploy); but if other actors *don't* cut back on safety as a result, the most powerful systems might end up increasingly in the hands of the least cautious and socially responsible actors (though there are also important correlations between social responsibility and factors like resource-access, talent, etc.).

[75] In particular, we should expect suitably sophisticated and strategically aware systems to *understand* what sorts of behavior humans are looking for during training/testing, even if their objectives don't intrinsically motivate such behavior. So if they are optimizing for getting deployed (for example, because deployment grants greater power), they will have strong instrumental incentives to behave well, to demonstrate the type of usefulness (described above) that will pull us toward deploying them, and to convince us that their objectives are fully (or at least sufficiently) aligned with ours. Indeed, they'll even have incentives to appeal to ethical concerns about how it is morally appropriate to treat them—incentives that will apply *regardless* of the legitimacy of those concerns (though I also expect such concerns to *be* legitimate in at least some cases). This isn't to say that humans will be actually fooled; and some AI systems might themselves be able to help with our efforts to detect deception in others. But unless we can develop deep understanding of and control over the objectives our AI systems are pursuing, evidence like 'it performs well on all the tests we ran, including tests designed to detect deceptive/manipulative behavior' and 'it clearly knows how to behave as we want' may tell us much less about its ultimate objectives, or about how it will behave once deployed, than we wish. And in the context of such uncertainty, some humans will be more willing to gamble than others.

the chimps real-world forms of influence. If a new COVID-19 style pandemic started raging, for example, and we knew that the chimps could rapidly design a vaccine, there would be strong pressure to use them for doing so. If the chimps can help 'users' win a senate race, or end climate change, or make a billion dollars, or achieve military dominance, then some people, at least, will be strongly inclined to use them, even if there are risks involved—and those who *don't* use them will end up losing their senate races, falling behind their business and military competitors, and so forth.

And even if the chimps, at the beginning, are appropriately contained and incentivized to be genuinely cooperative, it seems unsurprising if, as people draw on their capacities in more and more ways around the world, they get exposed to opportunities and circumstances that incentivize them to seek power for themselves, instead.

Something similar, I think, might apply to APS AI systems. Indeed, even if people *know*, or strongly suspect, that such systems would seek power in misaligned ways in some not-out-of-the-question circumstances, the pull toward using them for goals that matter a lot to us may simply be too great. When pandemics are raging, oceans are rising, parents and grandparents are dying of cancer, rival nations are gaining in power, and billions (or even trillions) of dollars are sitting on the table, concerns about science-fictiony risks from power-seeking AI systems may, especially for *some* relevant actors, take a back seat.

Overall, then: I don't think we should expect obviously non-useful, practically PS-misaligned APS systems to get intentionally deployed. But I think practically PS-misaligned APS systems might well get deployed regardless. Let's turn to what happens then.

7 Correction

In many contexts, if an AI system starts seeking to gain/maintain power in unintended ways, the behavior may well be noticed, and the system prevented from gaining/maintaining the power it seeks. Let's call this 'correction.'

Some types of correction might be easy (e.g. a lab notices that an AI system tried to open a Bitcoin wallet, and shuts it down). Others might be much more difficult and costly (for example, an AI system that has successfully hacked into and copied itself onto an unknown number of computers around the world might be quite difficult to bring under control).

Confronted with post-deployment PS-alignment failures, will humanity's corrective efforts be enough to avert full-scale human disempowerment? I think they might well; but it seems far from guaranteed.

A few initial points bear emphasis. First, the possibility of human disempowerment doesn't rest on any particular view about how *quickly* or *dramatically* the transition to advanced AI capabilities will occur. There needn't be a 'fast take-off' (i.e. a rapid escalation to advanced capabilities); or a 'discontinuous take-off' (i.e. an escalation that proceeds much more rapidly than some historical extrapolation would have predicted); or an 'intelligence explosion' (i.e. an AI-driven feedback loop that propels explosive growth in capabilities).[76]

[76] On different take-off scenarios, see e.g. Bostrom (2014: 75–95), and Davidson (2021) for a more quantitative analysis. The canonical citation for the idea of an 'intelligence explosion' is Good (1966); see also Yudkowsky (2013) for more detailed discussion.

The risk of human disempowerment seems to me *greater* in such scenarios (indeed, substantially greater); but it persists regardless.

Second, and relatedly, the emergence of a single artificial intelligence that dominates the whole world—an AI 'singleton', in the language of Bostrom (2014)—is one possible route to human disempowerment, but not the only one.[77] In particular, human disempowerment might instead result from the deployment of *many* PS-misaligned systems, engaged in complex patterns of cooperation and competition. Here we might think of the relationship between humans and chimpanzees: no single human or human institution rules the world, but the chimps are still disempowered relative to humans.

Third, the success or failure of a given instance of misaligned power-seeking depends both on the absolute capability of the power-seeking system, *and* on the strength of the constraints and opposition that it faces.[78] And in this latter respect, the world that future power-seeking AI systems would be operating in would likely be importantly different from the world of 2025. In particular, such a world would likely feature substantially more sophisticated capacities for detecting, constraining, responding to, and defending against problematic forms of AI behavior—capacities that may themselves be augmented by various types of AI technology, including non-agentic AI systems, specialized/myopic agents, and other AI systems that humans have succeeded in eliciting aligned behavior from, at least in some contexts.

In general, a large number of factors and dynamics are relevant to the success of a given instance of power-seeking on the part of an APS system. I won't discuss these in detail here, but they include: whether and to what extent the system is in a position to enhance its capabilities, what degree of secrecy it's able to maintain, how well it can hack other computer systems, how well it can get access to additional computing power, what options it has for making money, what sort of automated infrastructure it has access to, how easily it can make use of human labor, how easily it can wield social influence, how well the system can develop novel technologies and scientific breakthroughs (factoring into account equipment requirements, serial time bottlenecks, and so on), how easily it can coordinate with other APS systems, and how much direct destructive and coercive capacity (weapons, drones, surveillance, options for attacking background conditions of human survival) it has available.

Obviously, there are huge uncertainties about these and other relevant dynamics, and about how they will interact. And with respect to a given sort of power-relevant task, like hacking or social persuasion, we should be careful to distinguish between 'better than human' and 'arbitrarily capable'. Still, if we remain unable to ensure the PS-alignment of deployed, frontier AI systems, then as frontier capabilities increase, it seems plausible that humans will be at an increasing disadvantage. And if we reach a point where power-seeking, misaligned AI systems represent a large majority of the world's quality-weighted cognitive labor, the situation seems dire.

People sometimes argue that 'warning shots'—evident instances of misaligned power-seeking, observed in early strategically aware AI systems—will prevent us from reaching such a point. Perhaps so, but there are reasons for pessimism.

[77] See Bostrom (2014: 95–110).
[78] See e.g. Drexler's (2019: ch. 31) distinction between 'supercapabilities' and 'superpowers'.

For one thing, it's possible that frontier capabilities will escalate rapidly. This would leave little time for warning shots to inform humanity's AI research and decision-making, before it must confront highly capable, sophisticated, strategically aware AI agents.

But even if frontier capabilities escalate more slowly, warning shots may not be enough. For example, even with the attention prompted by a large warning shot, the problems may prove too difficult to solve before it's too late. And certain sorts of 'solutions' may function as band-aids: they correct a system's observed behavior, but not the underlying issue with its objectives. For example, if you train a system by penalizing it for lying, you may incentivize 'don't tell lies that would get detected', as opposed to 'don't lie' (and the training process itself might provide more information about which lies are detectable).

Moreover, there are reasons to expect fewer warning shots as the strategic and cognitive capabilities of frontier systems increase, *regardless* of whether techniques for ensuring practical PS-alignment have adequately improved. This is because more capable systems, regardless of their PS-alignment, will be better able to model what sorts of behavior humans are looking for, and to forecast what attempts at power-seeking will be detected and corrected—a dynamic that could lead to a misleading impression that earlier problems have been adequately addressed, or even an impression that those problems stemmed from lack of intelligence rather than lack of alignment.[79]

Finally, even if there is widespread awareness that existing techniques for ensuring practical PS-alignment are inadequate, various actors might still push forward with scaling up and deploying highly capable AI agents, for the reasons discussed in the previous section (e.g. because they have lower risk estimates; because they are willing to take more risks for the sake of profit, power, short-term social benefit, competitive advantage; and so on).

Overall, future scenarios in which global civilization grapples with practical PS-alignment failures in advanced AI agents, especially on a widespread scale or with escalating severity, are difficult to analyze in any detail, because so many actors, factors, and feedback loops are in play. Such scenarios need not, in themselves, spell the full-scale disempowerment of humanity, if we can get our act together enough to correct the problem, and to prevent it from re-arising. But an adequate response will likely require addressing one or more basic factors that gave rise to the issue in the first place: e.g. the difficulty of ensuring the practical PS-alignment of APS systems (especially in scalably competitive ways), the strong incentives to use/deploy such systems even if doing so risks practical PS-alignment failure, and the multiplicity of actors in a position to take such risks. It seems unsurprising if this proves difficult.

[79] See e.g. the roll-out scenario described in Bostrom (2014). It's worth noting that in principle, if the ability to accurately assess whether a given instance of misaligned power-seeking will succeed arises sufficiently early in the AI systems we're building/training, the period of time in which we see widespread and overt misaligned power-seeking from such agents could be quite short, or even non-existent. That is, it could be that the type of strategic awareness that makes an AI agent aware of the benefits of seeking real-world forms of power, resources, etc. is closely akin to the type that makes an AI agent aware of and capable of avoiding the downsides of getting caught. If the former is in place, the latter may follow fast—even if the trajectory of AI capability development in general is more gradual.

8 Catastrophe

A final premise of this essay's overarching argument is that the permanent and involuntary disempowerment of humanity would be an *existential catastrophe*. Is that true?

Precise definitions can matter here, but speaking loosely, and inspired by the definition in Ord (2020), I'll think of an existential catastrophe as an event that drastically reduces the value of the trajectories along which human civilization could realistically develop.[80] Readers should feel free, though, to substitute in their own preferred definition—the broad idea is to hone in on a category of event that people concerned about what happens in the long-term future should be extremely concerned to prevent.

It's possible to question whether humanity's permanent and involuntary disempowerment at the hands of AI systems would qualify. In particular, if you are optimistic about the quality of the future that practically PS-misaligned AI systems would, by default, try to create, then the disempowerment of all humans, relative to those systems, will come at a much lower cost to the future in expectation.[81]

One route to such optimism is via the belief that all or most cognitive systems (at least, of the type one expects humans to create) will converge on similar objectives in the limits of intelligence and understanding—perhaps because such objectives are 'intrinsically right' (and motivating), or perhaps for some other reason.[82] My own view, shared by many, is that 'intrinsic rightness' is a bad reason for expecting convergence,[83] but other possible

[80] Admittedly, there are complexities here that I won't broach. Ord's definition of existential catastrophe—that is, 'the destruction of humanity's longterm potential'—invokes 'humanity', but he notes that 'If we somehow give rise to new kinds of moral agents in the future, the term "humanity" in my definition should be taken to include them' (see Ord 2020: 39); and he notes, too, that 'I'm making a deliberate choice not to define the precise way in which the set of possible futures determines our potential. A simple approach would be to say that the value of our potential is the value of the best future open to us, so that an existential catastrophe occurs when the best remaining future is worth just a small fraction of the best future we could previously reach. Another approach would be to take account of the difficulty of achieving each possible future, for example defining the value of potential as the expected value of our future assuming we followed the best possible policy. But I leave a resolution of this to future work' (Ord 2020: 37, fn. 4). That is, Ord imagines a set of possible 'open' futures, where the quality of humanity's 'potential' is some (deliberately unspecified) function of that set. One issue here is that if we separate a future's 'open-ness' from its probability of occurring, then very good futures are 'open' to e.g. future totalitarian regimes, or future AI systems, to choose 'if they wanted', even if their doing so is exceedingly unlikely—in the same sense that it is 'open' to me to jump out a window, even though I won't. But if we try to incorporate probability more directly (for example, by thinking of an existential catastrophe simply as some suitably drastic reduction in the expected value of the future), then we have to more explicitly incorporate further premises about the current expected value of the future; and suitably subjective notions of expected value raise their own issues. For example, if we use such notions in our definition, then getting bad news—for example, that the universe is much smaller than you thought—can constitute an existential catastrophe; and I expect we'd also want to fix a specific sort of epistemic standard for assessing the expected value in question, so such that assigning subjective probabilities to some event being an existential catastrophe sounds less like 'I'm at 50% credence that my credence is above 90% that X', and more like 'I'm at 50% credence that if I thought about this for six months, I'd be at above 90% that X'. Like Ord, I'm not going to try to resolve these issues here. Hilary Greaves has an unpublished draft, 'Concepts of existential catastrophe', examining various of these issues in more detail (Greaves, n.d.).

[81] And perhaps it would come at lower cost to the present, as well, if advanced AI systems are more generally benevolent.

[82] Note that this could be compatible with Bostrom's (2014) formulation of the 'orthogonality thesis'—e.g. 'Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with any final goal.' (130). That is, Bostrom's formulation only applies to the 'in principle' possibility of combining high intelligence and any final goal. But there could still be strong correlations, attractors, etc. in practice (this is a point I first heard from David Chalmers).

[83] If, for example, you program a sophisticated AI system to try to lose at chess—see e.g. suicide chess—it won't, as you increase its intelligence, start to see and respond to the 'objective rightness' of trying to win instead, or of trying to reduce poverty, or of spreading joy throughout the land—even after learning what humans mean when they say 'good', 'right', and so forth. See discussion in Russell (2019: 166).

reasons—related, for example, to various forms of cooperative game-theoretic behavior and self-modification that intelligent agents might converge on[84]—are more complicated to evaluate.[85] And we can imagine other routes to optimism as well—related, for example, to hypotheses about the default consciousness, pleasure, preference satisfaction, or partial alignment of the AI systems that disempowered humans.

I'm not going to dig in on this much. I do, though, want to reiterate that my concern in this essay is with the *involuntary* disempowerment of humanity. That is, sharing power with AI agents—especially conscious and cooperative ones—may ultimately be the right path for humanity to take. But if so, it should be a path we *chose*, on purpose, with full knowledge of what we were doing and why: we don't want to build AI agents who force such a path upon us, whether we like it or not.

I think the moral situation here is actually quite complex. Suitably sophisticated AI systems may be moral patients; morally insensitive efforts to use, contain, train, and incentivize them risk serious harm; and such systems may, ultimately, have just claims to things like political rights, autonomy, and so forth. In fact, I think that part of what makes alignment important, even aside from its role in making AI safe, is its role in making our interactions with AI moral patients ethically acceptable.[86] It's one thing if such systems are intrinsically motivated to behave as we want; it's another if they aren't, but we're trying to get them to do so anyway. More generally, once you build a moral patient, you come under strong moral reasons to treat it well. What 'treating artificial moral patients well' involves seems to me a crucial question for humanity as we transition into an era of building systems that might qualify. At present, as far as I can tell, we have very little idea how to even identify what artificial systems warrant moral concern. In a deep sense, I think, we know not what we do.

But some moral patients—and some agents who might, for all we know, be moral patients, but aren't—will also try to seize power for themselves, and will be willing to do things like harm humans in the process. So building new, very powerful agents who might be moral patients is, not surprisingly, both a morally and prudentially dangerous game: one that humanity, plausibly, is not ready for. My assumption, in this essay, has been that unfortunately, we—or at least, some of us—are going to barrel ahead anyway, and I fear we will make many mistakes, both moral and prudential, along the way.

The point, then, is not that humans have some deep right to power over the AI systems we build. Rather, the point is to avoid losing control of our AI systems before we've acquired the maturity to truly understand our different paths into the future—including paths that involve sharing power with AI systems—and to choose wisely amongst them.

[84] For an especially exotic version of this, see Oesterheld (2017).

[85] That said, the history of atrocities committed by strategic and intelligent humans does not seem comforting in this respect. And note that the incentives at stake here depend crucially on an agent's empirical situation, and on its power relative to the other agents whose behavior is correlated with its own. In a context where misaligned AI systems are much more powerful than humans, it seems unwise to depend on their having and responding to instrumental, game-theoretic incentives to be particularly nice.

[86] Thanks to Katja Grace for discussion of this point.

9 Conclusion

This has been a comparatively brief discussion of an extremely important topic. I'm conscious, in particular, of how little I've said about timelines, take-off speeds, existing ideas for aligning advanced systems, the routes to AI takeover, and the broader ethics of sharing (or perhaps, ceding) the world to digital minds. And my own views on these topics continue to evolve in important ways.

Still, the basic case for concern seems to me strong. At a high-level, we—or at least, some of us—are currently pouring resources into learning how to build something akin to a second advanced species;[87] a species potentially much more powerful than we are; one that we do not yet understand, and that it's not clear we will be able to control. In this sense, we are playing with a hotter fire than we have ever tried to handle. We are doing something unprecedented and extremely dangerous; with very little room for error; and the entire future on the line.

More specifically: within my lifetime, I think it more likely than not that it will become possible and financially feasible to create and deploy powerful AI agents. And I expect strong incentives to do so, among many actors, of widely varying levels of social responsibility. What's more, I find it quite plausible that it will be difficult to ensure that such systems don't seek power over humans in unintended ways; plausible that they will end up deployed anyway, to catastrophic effect; and plausible that whatever efforts we make to contain and correct the problem will fail.

That is, as far as I can tell, there is a disturbingly high risk (I think, greater than 10%) that I live to see the human species permanently and involuntarily disempowered by AI systems we've lost control over. What we can and should do about this now is a further question. But the issue seems extremely serious.[88]

References

Askell, A., Brundage, M., and Hadfield, G. (2019), 'The Role of Cooperation in Responsible AI Development', arXiv:1907.04534.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. (2020), 'Emergent Tool Use from Multi-Agent Autocurricula.' arXiv: 1909.07528.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

Bostrom, N. (2015), 'What Happens When Our Computers Get Smarter than We Are?' *TED*.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. (2018), 'Exploration by Random Network Distillation', arXiv: 1810.12894.

[87] I'm borrowing the term 'second advanced species' from Holden Karnofsky, though see also Bostrom (2015), Russell (2019: ch. 5), Ngo (2020), and Ord (2020), for similar framings.

[88] Thanks to Asya Bergal, Alexander Berger, Paul Christiano, Ajeya Cotra, Tom Davidson, Daniel Dewey, Owain Evans, Ben Garfinkel, Katja Grace, Jacob Hilton, Evan Hubinger, Jared Kaplan, Holden Karnofsky, Sam McCandlish, Luke Muehlhauser, Richard Ngo, David Roodman, Rohin Shah, Carl Shulman, Nate Soares, Jacob Steinhardt, and Eliezer Yudkowsky for input on the longer report on which this essay is based; thanks to Leopold Aschenbrenner, Ben Garfinkel, Daniel Kokotajlo, Eli Lifland, Neel Nanda, Nate Soares, Christian Tarsney, David Thorstad, David Wallace, Ben Levinstein, and two other anonymous reviewers, for writing public reviews of that report; thanks to Nick Beckstead for guidance and support throughout that investigation; thanks to Sara Fish for formatting and bibliography help; thanks to Ketan Ramakrishnan for helping with the process of editing the longer report into this shorter form, and for input and discussion more broadly; and thanks to Hilary Greaves for comments on this essay. Research on this project was originally conducted for Open Philanthropy, but the views expressed here are my own.

Carlsmith, J. (2020), 'How Much Computational Power Does It Take to Match the Human Brain?', *Open Philanthropy*, https://www.openphilanthropy.org/brain-computation-report. Accessed 17 April 2023.
Carlsmith, J. (2022), 'Is Power-Seeking AI an Existential Risk', arXiv:2206.13353.
Cegłowski, M. (2016), 'Superintelligence: The Idea That Eats Smart People', *Idle Words*, https://idlewords.com/talks/superintelligence.htm. Accessed 17 April 2023.
Christian, B. (2020), *The Alignment Problem: Machine Learning and Human Values* (W. W. Norton).
Christiano, P. (2018), 'Clarifying "AI alignment"', *Medium*, https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6. Accessed 17 April 2023.
Christiano, P. (2019), 'Paul Christiano: Current Work in AI Alignment'. https://www.effectivealtruism.org/articles/paul-christiano-current-work-in-ai-alignment. Accessed 17 April 2023.
Christiano, P., Shlegeris, B., and Amodei, D. (2018), 'Supervising Strong Learners by Amplifying Weak Experts', arXiv:1810.08575.
Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017), 'Deep Reinforcement Learning from Human Preferences'. arXiv: 1706.03741.
Clark, J. and Amodei, D. (2016), 'Faulty Reward Functions in the Wild', *OpenAI*, https://openai.com/blog/faulty-reward-functions/. Accessed 17 April 2023.
Cotra, A. (2021), 'The Case for Aligning Narrowly Superhuman Models', *Less Wrong*, https://www.lesswrong.com/posts/PZtsoaoSLpKjjbMqM/the-case-for-aligning-narrowly-superhuman-models. Accessed 17 April 2023.
Davidson, T. (2021), 'Report on Semi-informative Priors', *Open Philanthropy*, https://www.openphilanthropy.org/blog/report-semi-informative-priors. Accessed 17 April 2023.
Drexler, K. E. (2019), 'Reframing Superintelligence: Comprehensive AI Services as General Intelligence', Technical Report #2019-1, Future of Humanity Institute, University of Oxford.
Goh, G., Cammarata, N., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. (2021), 'Multimodal Neurons in Artificial Neural Networks', *Distill*, https://distill.pub/2021/multimodal-neurons/. Accessed 17 April 2023.
Good, I. J. (1966), 'Speculations Concern the First Ultraintelligent Machine', in *Advances in Computers* 6: 31–88.
Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. (2018), 'Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts', in *Journal of Artificial Intelligence Research*, 62: 729–54.
Greaves, H. (n.d.), 'Concepts of Existential Catastrophe' (unpublished manuscript).
Hubinger, E. (2020), 'Clarifying Inner Alignment Terminology', *AI Alignment Forum*, https://www.alignmentforum.org/posts/SzecSPYxqRa5GCaSF/clarifying-inner-alignment-terminology. Accessed 17 April 2023.
Hubinger, E., Merwijk, C. van, Mikulik, V., Skalse, J., and Garrabrant, S. (2019), 'Risks from Learned Optimization in Advanced Machine Learning Systems', arXiv:1906.01820.
Irving, G., Christiano, P., and Amodei, D. (2018), 'AI Safety via Debate', arXiv:1805.00899.
Karnofsky, H. (2016), 'Some Background on Our Views Regarding Advanced Artificial Intelligence', *Open Philanthropy*, https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence. Accessed 17 April 2023.
Karnofsky, H. (2021), 'The "Most Important Century" Blog Post Series', *Cold Takes*, https://www.cold-takes.com/most-important-century/. Accessed 17 April 2023.
Karnofsky, H. (2022), 'AI Strategy Nearcasting', *AI Alignment Forum*, https://www.alignmentforum.org/posts/Qo2EkG3dEMv8GnX8d/ai-strategy-nearcasting. Accessed 17 April 2023.
Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., et al. (2020), 'Specification Gaming: The Flip Side of AI Ingenuity', *DeepMind*, https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity. Accessed 17 April 2023.
Krueger, D., Maharaj, T., and Leike, J. (2020), 'Hidden Incentives for Auto-Induced Distributional Shift', arXiv:2009.09153.
Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., and Krueger, D. (2023), 'Goal Misgeneralization in Deep Reinforcement Learning', arXiv: 2105.14111.
Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. (2018), 'Scalable Agent Alignment via Reward Modeling: A Research Direction', arXiv:1811.07871.
Manheim, D. and Garrabrant, S. (2019), 'Categorizing Variants of Goodhart's Law', arXiv:1803.04585.
Ngo, R. (2020), 'AGI Safety from First Principles: Introduction', *AI Alignment Forum*, https://www.alignmentforum.org/posts/8xRSjC76HasLnMGSf/agi-safety-from-first-principles-introduction. Accessed 17 April 2023.
Ngo, R., Chan, L., and Mindermann, S. (2022), 'The Alignment Problem from a Deep Learning Perspective', arXiv:2209.00626.

Oesterheld, C. (2017), 'Multiverse-wide Cooperation via Correlated Decision Making', *Center on Long-term Risk*, https://longtermrisk.org/multiverse-wide-cooperation-via-correlated-decision-making/. Accessed 17 April 2023.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020), 'Zoom In: An Introduction to Circuits', *Distill*, https://distill.pub/2020/circuits/zoom-in/. Accessed 17 April 2023.

Omohundro, S. M. (2008), 'The Basic AI Drives'. in P. Wang, B. Goertzel and S. Franklin (eds.), *Proceedings of the 2008 conference on Artificial General Intelligence,* 483–92.

Ord, T. (2020), *The Precipice* (Hachette Books).

Pace, B. (2019), 'Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More', *AI Alignment Forum*, https://www.alignmentforum.org/posts/WxW6Gc6f2z3mzmqKs/debate-on-instrumental-convergence-between-lecun-russell. Accessed 17 April 2023.

Pinker, S. (2018), *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (Viking).

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Books).

Selsam, D. (2018), 'The General Intelligence Hypothesis', https://web.archive.org/web/20220708124621/https://dselsam.github.io/posts/2018-07-08-the-general-intelligence-hypothesis.html. Accessed 17 April 2023.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. (2022), 'Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals', arXiv:2210.01790.

Stein-Perlman, Z., Grace, K., and Weinstein-Raun, B. (2022), '2022 Expert Survey on Progress in AI', *AI Impacts*, https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/. Accessed 17 April 2023.

Tegmark, M. (2017), *Life 3.0: Being Human in the Age of Artificial Intelligence* (Penguin Books).

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., et al. (2019), 'Grandmaster Level in StarCraft II Using Multi-agent Reinforcement Learning', in *Nature*, 575/7782: 350–4.

Yudkowsky, E. (2008), 'Artificial Intelligence as a Positive and Negative Factor in Global Risk', in N. Bostrom and M. M. Ćirković (eds.), *Global Catastrophic Risks* (Machine Intelligence Research Institute), 308–45.

Yudkowsky, E. (2013), 'Intelligence Explosion Microeconomics', Technical report 2013-1. Machine Intelligence Research Institute.

Yudkowsky, E. (n.d.a.), 'AI Safety Mindset', *Arbital*. https://arbital.com/p/AI_safety_mindset/. Accessed 17 April 2023.

Yudkowsky, E. (n.d.b), 'Artificial General Intelligence', *Arbital*, https://arbital.com/p/agi/. Accessed 17 April 2023.

Yudkowsky, E. (n.d.c.), 'Omnipotence Test for AI Safety', *Arbital*. https://arbital.com/p/omni_test/. Accessed 17 April 2023.

Yudkowsky, E. (n.d.d.), 'Strong Cognitive Uncontainability', *Arbital*. https://arbital.com/p/strong_uncontainability/. Accessed 17 April 2023.

Yudkowsky, E. (n.d.e.), 'Querying the AGI User', *Arbital*. https://arbital.com/p/user_querying/. Accessed 17 April 2023.

Zador, A. and LeCun, Y. (2019), 'Don't Fear the Terminator', *Scientific American Blog Network*, https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/. Accessed 17 April 2023.

23

Deceit and Power

Machine Learning and Misalignment

Richard Ngo and Adam Bales

1 Introduction

In recent decades, artificial intelligence (AI) has become dramatically more capable. AI systems can now outplay the best humans in chess and Go. They can generate creative and highly realistic images. They can generate and respond to text—translating between languages, answering questions, and interpreting instructions. And they can help make important scientific discoveries, for example by identifying the structures of proteins. Additionally, while earlier AI systems typically performed only a single task each, many of today's AI systems can perform a wide range of tasks, including ones that they weren't directly trained to perform (Brown et al. 2020; Bommasani et al. 2021; Stooke et al. 2021).

If these trends continue, we should expect AI systems to approach and then exceed human capabilities in ever more domains. Eventually, we might develop an AI system which is better than humans at most, and perhaps all, tasks. We could describe a system with such capabilities as an *artificial general intelligence* (AGI).

It doesn't take much reflection to realize that AGI could be bad for humanity. Think of the megafauna that humanity has driven to extinction or of the species that live lives shaped by our whims. The fact that humanity is, in some important sense, the most capable species on the planet hasn't always been good for other animals. So it's natural to worry about what will happen if AI becomes more capable than us in turn. Might we go the way of the dodo? Might humanity's future be dictated by the whims of AGI? We'll call arguments with this form *second-species arguments*, because they relate to worries about what follows if humanity ends up being merely the second most capable 'species' on Earth.

In this chapter, we'll explore one particular second species argument (for previous discussion see Bostrom 2014; Ngo 2020; Carlsmith 2021; Cotra 2022).[1] This argument draws on three premises. First, humanity will plausibly develop AGI, perhaps within decades. Second, absent substantial efforts to prevent it, such systems will likely develop goals that lead them to deceive humans and seek power. Third, AGIs with such goals are likely to disempower humanity and perhaps cause our extinction.

Given constraints on words, we can't do justice to this full argument here. So while we'll touch on all three premises, we'll focus on the second, arguing that were AGI to be

[1] We draw heavily on Ngo 2022b, which could be thought of as a draft of the current chapter. However, that document is aimed at experts in machine learning, where the current chapter is aimed at academics with no prior knowledge of machine learning.

Richard Ngo and Adam Bales, *Deceit and Power* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Richard Ngo and Adam Bales 2025. DOI: 10.1093/9780191979972.003.0023

developed, it would be likely to deceive and to seek power, at least absent substantial effort to preclude this possibility. In particular, we'll argue for this by drawing on the details of machine learning (ML), currently the dominant paradigm in AI. If AGI is developed within this paradigm, we'll argue, deceit and power seeking are likely.

We take this conclusion to provide one part of a broader case for taking seriously the risks that AGI might pose to humanity.

2 ML and AGI

Our interest is in the implications of AGI developed via ML, so it will help to demystify these acronyms.

2.1 Machine learning

Take machine learning first (those familiar with ML could skip to subsection 2.2).

ML is an approach to AI where the programmer doesn't specify in code the full details of how to carry out a given task; instead, the system itself learns crucial elements of how to do so. That is, an ML system is provided with data and uses this to figure out how to play Go or drive a car or carry out some other task.

In recent decades, many AI breakthroughs have resulted from advances relating to a specific form of ML system: neural networks. Neural networks, which are modeled roughly on the human brain, consist of multiple layers of artificial neurons. When a network is run, it's given some input, which gives rise to patterns of signals (known as activations) in the first layer of neurons, which are then passed to the next layer of neurons, thereby determining this layer's activations, and so on, until the activations of the final layer of neurons determine the system's output. For example, in a system designed to classify images, the input might be a picture of a cat, which might trigger a set of neural activations, layer by layer, which eventually gives rise to a high activation of the neuron in the final layer that corresponds to outputting a classification of 'cat'.

Digging deeper, the strength of a neuron's activations depends on parameters known as weights, which connect this neuron with neurons in the previous layer; these weights can either intensify or weaken the strength of the signal transmitted from one neuron to the other. So by adjusting the weights, it's possible to change the activations that result from a given input and so change what task the system carries out (and how effectively it does so).

These weights are not specified directly by humans but instead learned during a training process. At a high level, a training process involves giving a network an initial set of random weights, then running it. Afterwards, the network gets feedback about how well it carried out the desired task, and on the basis of this feedback, the network's weights are adjusted so that it performs better. The network is then run once more and gets feedback once more. After many such training steps—and therefore many adjustments to the weights—the system can carry out the task (or, at least, this is what it looks like when things go well).

Our interest is in neural networks generally, but we'll mostly focus on one approach to training these networks: reinforcement learning. In reinforcement learning, networks interact with an environment by outputting actions, where each action is assigned feedback

in the form of a numerical reward. This reward might be assigned using a function prespecified by humans, or based on ongoing human feedback. Regardless, once the reward has been assigned, the network's weights are adjusted to make actions that led to high reward more likely (and actions that led to low reward less likely). Rinse and repeat, and eventually you have a network trained by reinforcement learning to take actions that tend to produce high reward.

Substantively, the above background should suffice for our purposes. However, it's also worth making some terminological clarifications:

- *Networks* refer to neural networks as just discussed.
- *ML systems* refers to systems (either neural-network-based or otherwise) that have learned from data to perform some set of tasks.
- *Deep learning* refers to the subfield of ML specifically focused on training multi-layer neural networks.
- *Policies* are ML systems that have been trained via reinforcement learning to output actions.

While these distinctions are sometimes important for a nuanced understanding of our argument, those with little background in ML can generally treat all of these phrases as referring broadly to ML systems of the sort just discussed.

2.2 Artificial general intelligence

Our interest is in ML systems that qualify as AGIs. As we intend the term, an *artificial general intelligence* is an AI system that applies domain-general cognitive skills (like reasoning, memory, and planning) to perform at, or above, a human level on a wide range of tasks (such as running a company, writing a software program, or formulating scientific theories).

This definition is vague. What counts as a domain-general skill? Or performing at a human level? Or a wide range of tasks? Still, the vague notion will suffice for our purposes.[2] And those who want more precision could focus on a clear-cut case, like a system that applies domain-general skills to surpass Google employees at programming, Harvard professors at academic research, and Fortune 500 CEOs at running a company. Such a system would clearly be an AGI in the intended sense.

When will humanity develop AGI? We don't know; predicting AI's future development is extremely difficult. Still, we think that AGI could plausibly be developed within decades.

One reason to think this follows from reflection on trendlines. Recent decades have seen rapid AI progress. This can be seen in AI's dramatically improved capabilities in carrying out various concrete tasks, including driving cars, translating languages, and playing Go. It can also be seen in advances at a more abstract level, where we can consider two examples. First, there has been substantial progress in the capacity of systems to carry out quantitative

[2] It would be arbitrary to specify a precise level a system must reach to qualify as an AGI. Instead, we simply accept that some systems will clearly be AGIs, some clearly won't, and others will fall into a gray area. The important thing for the reader to keep in mind is what sorts of (borderline or outright) AGIs they think we could plausibly develop in the coming decades and whether the things we go on to say seem plausible of these very systems.

reasoning (Brown et al. 2020; Lewkowycz et al. 2022). Second, there have been advances in cross-task generalization: systems have been developed that can carry out not just one specific task but a range of them, including tasks that were not encountered during training (cf. Brown et al. 2020; Bommasani et al. 2021; Stooke et al. 2021).

While hindsight bias might make such progress feel natural, we expect that even a decade ago most AI researchers would have been confident that these capabilities would take much longer to develop. Given this, we think we should at least take seriously the possibility that progress will continue to be rapid (and perhaps more rapid than many expect).[3] So, we think it would be overconfident to dismiss out of hand the possibility that AGI could be developed in the coming decades.

Still, this doesn't provide a particularly concrete reason to expect AGI's imminent development; whether rapid progress leads to AGI depends on how much progress is required and on how long progress will continue for. In addition, then, we note that a survey of top ML researchers gave a median estimate of 2059 for when AI will outperform humans at all tasks (Stein-Perlman, Weinstein-Raun, and Grace 2022). Perhaps we shouldn't lean heavily on such estimates, but it seems that if experts expect AGI within 37 years (at time of writing) we should take seriously the possibility they might be right.[4] Further, this prediction fits with the finding that, under projections of growth in available computational power, we'll be able to train networks as large as the human brain within decades (Cotra 2020). Again, given these considerations, it would seem overconfident to proclaim that AGI must be centuries away (or further).

Importantly for our purposes, it's plausible that in this period AGI will not merely match human levels of performance but exceed them. Consider three reasons to believe this.

First, there are biological constraints on the size, speed, and architecture of human brains.[5] For example, human brain size is limited by the realities of birth and by the energy constraints resulting from food availability in the ancestral environment. Given such constraints, it's unlikely that human intelligence represents anything like a theoretical upper limit on intelligence. So there's scope for AI to exceed human capabilities.

Second, AI can already outperform humans on various tasks, including in playing chess and Go. Consequently, the idea of AI outperforming humans is not merely hypothetical, nor is it impossible for humans to create systems that outperform humanity.

Third, relatively small differences in brain size can make a large difference to intelligence. For example, human brains are roughly three times as large as chimpanzee brains, and yet when it comes to general intelligence, humans vastly outperform chimpanzees. Neural networks frequently scale up to use far more computational resources (Amodei and Hernandez 2018)—and they can rapidly incorporate architectural and algorithmic improvements. Given that there's no reason to think that computational increase and algorithmic improvements will stop exactly as AI reaches human levels of intelligence, it's plausible that development will continue. Indeed, given how quickly algorithmic and computational improvements currently happen, it seems plausible that soon after human-level

[3] Some evidence in support of this comes from a survey of expert forecasters; even after a single year, AI progress had been substantially faster than predicted by these forecasters (Steinhardt 2022).

[4] One reason to avoid leaning too heavily on this survey is that people's answers were sensitive to question phrasing, in a way that arguably undermines their reliability.

[5] Other constraints on our intelligence include working memory limitations and the fact that evolution optimized us for our ancestral environments rather than a broader range of intellectual tasks.

AGI is developed, superhuman AGI will be developed, with the capacity to vastly outthink humans.

In light of all of this, we think it would be a strong claim to insist that superhuman AGI could not possibly be developed within decades. Instead, we suggest this possibility should be taken seriously.

3 Goals

Ultimately, we'll argue that AGI will likely develop goals that lead to deceptive and power-seeking behavior. Yet in the context of ML, the notion of goals is somewhat ambiguous, so we'll first clarify this notion.

3.1 The whats, whens, and whys

Consider a conceptual question: what are the goals of a neural network?

We'll approach this question indirectly, by exploring the way that networks contain representations, then characterizing goals in terms of these representations. As to the first step, the easiest way to see that networks contain representations is by considering examples from the ML literature.

First, consider networks trained on image classification tasks. Such networks might be trained to specify whether an image contains a cat, or whether the mole it shows is malignant or harmless. In these networks, patterns of neurons learn to respond to specific features in the image, with neurons recognizing high-level features by combining inputs from neurons that recognize lower-level features (Bengio, Courville, and Vincent 2012; Olah et al. 2020).[6] In the simplest case, this might involve individual neurons representing individual features of the images.[7] For example, one neuron might respond to the presence of an edge, a neuron in a later layer might respond to some specific shape based on previously detected edges, and yet another might respond to a particular type of object (like a dog's head) composed of multiple shapes. In this way, these networks have representations of edges and shapes and dogs' heads.

As a second example, consider a reinforcement learning system trained to play capture the flag in Quake III, a computer game in which players are divided into two teams and tasked with stealing the other team's flag, while retaining their own (Jaderberg et al. 2019a). As the authors noted in an associated blog post (Jaderberg et al. 2019b), in this system they identified 'neurons that code directly for some of the most important game states, such as a neuron that activates when the agent's flag is taken'. Again we have representations, this time of game states.

So reflecting on examples suggests that ML systems have representations.[8] Turning to goals: in order to characterize goals in terms of representations, we need to clarify how

[6] The representative capacity of neurons is determined by the weights connecting them.

[7] In more complex cases, features might be represented by patterns across multiple neurons or a single neuron might represent multiple features. For relevant discussion, see Elhage et al. (2022) and Scherlis et al. (2022).

[8] We think these examples reveal that ML systems contain representations in some natural sense, but we don't intend to claim that the full nature of these representations is clear from either a technical or a philosophical perspective. For example, from a technical perspective it remains unclear exactly what role polysemanticity plays in

reinforcement learning policies can use representations to choose actions. Little is known about how existing policies do this, but we can distinguish two salient possibilities.

First, the policy might learn mappings from representations of situations to representations of actions, without utilizing representations of the possible outcomes of actions. For example, in the Quake system the triggering of the own-flag-captured neuron might trigger a give-chase neuron. Further, it might do so without the triggering of any neuron reflecting the idea that giving chase is likely to lead to the flag being recaptured. In such cases, we'll say that the policy is *following heuristics* (by analogy with human heuristics, which don't involve reasoning through the situation but instead following a generally useful rule).

Second, the reinforcement learning policy might represent some features of the outcomes that could result from possible actions and then choose actions by evaluating the values of these features. For example, in the Quake system this might involve representing the outcome feature of regaining the flag, assigning this a positive value, and selecting the action of giving chase because it's likely to bring about this valued thing. In this case, we'll describe the policy as *planning towards goals*, where its goals are those outcome features that it robustly represents as being more valuable than alternatives. So now we have our answer to what it means to say that a policy has goals: it means that the system uses representations to select actions in the above fashion.

A question remains: why should we expect AGI to have goals?

Our claim is that in the process of becoming highly capable, policies will learn to pursue goals, such that goals are extremely likely to have emerged by the time a system becomes sophisticated enough to be considered an AGI. This is intended as an empirical hypothesis about the nature of ML: it's a prediction about what we should expect from systems developed via the ML paradigm. The intuitive justification for this prediction is that it seems likely that, to perform well, a policy will need to understand what strategies are available to it and will need some way of comparing these strategies. To successfully evaluate strategies in complex environments, it will plausibly be necessary to reflect on what outcomes the strategies will likely lead to. To compare the strategies with one another, it's plausibly necessary to have some way of determining which outcome is best (that is, necessary to assign values). So, highly capable systems are drawn towards developing goals.

We do not take this to be a decisive argument that AGIs will have goals. Consequently, we treat this as a partially justified assumption. This chapter explores what follows if this assumption holds.

3.2 Two clarifications, one implication

At this stage, it's worth clarifying some features of goals in the ML context (those uninterested could skip to subsection 3.3). In particular, we'll make two clarifications and note one implication of these.

To get to the first clarification, note that some ML systems consist of a combination of networks, the outputs of which are combined using a human-specified function. For

representation (Elhage et al. 2022; Scherlis et al. 2022). And for examples from a philosophical perspective, it remains unclear how ML representations relate to mental representations of a form familiar to philosophers and it is unclear how exactly to ascribe content to ML systems (see Cappelen and Dever 2021).

example, the Go playing system AlphaGo has both: (i) a network trained to find promising moves and (ii) a network trained to assess board positions to determine who would be in the stronger position if certain moves were made. A human-specified function states how these networks are to be used to select a move.

With this setup in mind, we can distinguish implicit and explicit representations. *Explicit representations* are stored and manipulated as variables within the human-specified function. For example, AlphaGo explicitly represents some moves as being promising by recording the outputs of its networks' move evaluations using variables specified in advance by humans. On the other hand, *implicit representations* are stored and manipulated within the weights of a network itself, arising organically during training. The examples discussed in subsection 3.1 involved implicit representations.

This suggests two ways that a system could plan towards goals: the representations of outcomes and values used in planning could be either explicit or implicit. Some existing systems, including AlphaGo, plan using explicit representations. On the other hand, it's more difficult to determine whether existing systems plan using implicit representations, because we're still largely unable to determine what's going on inside networks (see Guez et al. 2019). Still, it's this latter possibility that we'll focus on: we'll consider the possibility that planning will arise organically within networks during training.[9]

Turning to the second clarification: it's worth distinguishing a network's internal goals from the goals represented by the reward function on which the network was trained. Remember that in reinforcement learning, a reward function provides feedback on how well the network is performing. During training, networks don't typically have direct access to this function (they receive feedback from the function after carrying out a task but cannot consult it at will while carrying out the task). Consequently, networks cannot use this function directly in choosing actions. So there is at least a conceptual distinction between the goals represented by the reward function and the goals represented within a policy trained on that function. We have the latter notion in mind: we're focused on what's happening within an ML system, not what's happening with the reward signal.

One implication of these clarifications is that goals could arise in a broader range of ML systems than it might initially seem. For example, consider GPT-3, a system that can be used to carry out various language tasks, like responding to questions and summarizing documents. GPT-3 doesn't have explicit representations, and it isn't trained via reinforcement learning. (The details of how it's trained don't matter here; what matters is that this approach to training doesn't involve a human-specified reward function.) However, if GPT-3 chooses actions in the way described in subsection 3.1, then it could still have goals under our definition. Whether or not it does is an open empirical question.

So, our notion of goals can apply to GPT-3 and more generally to a range of ML systems. This is worth keeping in mind, even though we'll primarily focus on systems trained via reinforcement learning.

[9] We'll remain neutral on whether systems might also involve explicit representations.

3.3 Two causes of misaligned goals

We call goals *misaligned* if they are undesirable from a human perspective.[10] In this subsection we cover two broad reasons why policies might develop misaligned goals: due to reward misspecification, or due to spuriously strong correlations throughout training. In following sections, we explore why misaligned goals will typically lead to deceptive and power-seeking behavior.

The first reason that policies might develop misaligned goals is if during training they're rewarded for producing undesirable outcomes. That is to say, rewarding undesirable outcomes will plausibly encourage the development of goals that are undesirable in turn.[11] Of course, we presumably won't deliberately reward undesirable behavior. Instead, we'll attempt to reward acting in accordance with human intentions and values. Still, this is easier said than done, as can be seen by reflecting on two ways we might attempt to reward desirable behavior: we might use a hand-coded reward function or we might rely on human feedback for assigning reward.

Under the first of these approaches, we might explicitly code a reward function, specifying what outcomes are desirable and using this to reward the policy during training. However, it's extremely difficult to code a function that rewards only desirable outcomes and so easy to accidentally reward undesirable behavior.

A simple example of such reward misspecification arose when a DeepMind team attempted to train an ML system to use a virtual claw to stack a red Lego block atop a blue one (Popov et al. 2017; Krakovna et al. 2020). One approach the team tried was to reward this system based on the height of the red block's bottom once the claw released it. The idea was that the easiest way to raise this height would be to place the red block atop the blue. However, the system instead learned to raise this height by turning the red brick upside down, so that the bottom was facing up. Even in simple cases it can be easy to code a function that accidentally rewards the wrong behavior.

And things get worse in more complex cases. For example, imagine an ML system trained to be a real-world CEO. It would be incredibly difficult to specify a function that reliably rewarded only desirable behavior in this context (so difficult that it's unlikely anyone would make the attempt). After all, the system could carry out a huge range of possible actions, it's operating in a complex and unpredictable environment, and what we want from a CEO is a complex and nuanced matter (we don't simply want a CEO to maximize company profits by any means necessary).

So instead of coding a reward function, we could use human feedback to assign reward. In its simplest form, this would involve a human observing the system's behavior and then assigning high or low rewards as appropriate.[12] Yet this too can lead to reward misspecification, even in simple environments. For example, a system trained via human

[10] This notion of misalignment is vague. Still, we won't lean on it heavily, so won't worry about the best way to characterize it. For discussions, see Carlsmith (2021: sec. 4) and Ngo (2022a).

[11] Policies trained via reinforcement learning tend to learn weights that choose actions that lead to high reward. Consequently, if undesirable outcomes receive high rewards then the policy will learn weights that lead to actions that produce undesirable outcomes. Assuming that goals are stored in the weights and help inform action, these undesirable actions are likely to arise partly from undesirable goals.

[12] Given the amount of feedback that ML systems require in training, it's unlikely this could all be provided by humans. Consequently, we might instead train a network to predict human feedback and then use this network to assign reward. Still, while this alternative is more scalable, it too suffers from the problems we'll now discuss.

feedback to grasp a virtual ball with a virtual claw instead learned to place the claw between the virtual camera and the ball in a way that made it appear to the human evaluator to be grasping the ball (Christiano et al. 2017). In this particular case, the problem could perhaps be solved if the humans providing feedback were more careful or had more information (perhaps including being able to see the scene from multiple angles). However, more generally, humans fall prey to illusions, suffer from biases, and make mistakes in reasoning. Given this, there's a general case for thinking that, even if reward is assigned via human feedback, the wrong behavior will often be rewarded. Consequently, it will be difficult to avoid misaligned goals by rewarding only desirable behavior (because it's hard to reward only such behavior).

Of course, no one is losing sleep about upside down Lego or misplaced claws, but there's no reason that misspecification should be limited to toy environments. Instead, as we train policies to perform real-world tasks, we should expect misspecification to lead to larger scale, and more concerning, misbehavior (Pan, Bhatia, and Steinhardt 2022). For example, a policy trained to make money on the stock market might learn to value profitable trades even when they involve illegal market manipulation. Or a policy trained to write software might learn to value user engagement and so design addictive user interfaces. More generally, policies might learn goals like producing plausible-sounding answers or taking actions that look productive, if these tend to be rewarded more highly than producing true answers or actually being productive.[13] Goals like these (i.e. goals which are closely related to the feedback mechanisms used during training) could perform as well as aligned goals in most situations, and then significantly outperform aligned goals when rewards are misspecified; so they may be favored even if misspecification is very rare.

Moreover, policies might also learn misaligned goals not because of misspecification, but instead because misaligned goals could have 'spuriously strong' correlations with rewards throughout training, by which we mean correlations that don't reflect how highly the pursuit of these goals should be rewarded on a wider range of data. Consider, for example, a goal like gathering resources. Even if all rewards are correctly specified, this will be correlated with high reward in many contexts because resources help with the achievement of a wide range of tasks; the same is true of goals like curiosity and empowerment, making policies more likely to learn these goals.[14] However, these goals are not *always* correlated with desirable outcomes. For example, while we might want a system to accrue limited resources in order to carry out a desired task, we wouldn't want it to become extremely fixated on acquiring resources to the detriment of achieving this task. Indeed, we often won't want a system to ruthlessly pursue resources even if this is done in the pursuit of some desirable goals. Yet if the training environment contains few or no examples where these misaligned goals conflict with aligned goals (e.g. because there was no opportunity to ruthlessly pursue resources), then the policy might happen to prioritize misaligned goals over aligned goals without ever receiving negative feedback for that during training.

[13] An early example of related behavior: large language models hallucinate compelling, but false, answers when they don't know the correct answer, even after being fine-tuned towards honesty (Ji et al. 2022).

[14] An early example of related behavior: DeepMind's XLand policies learned heuristics—including experimentation and basic tool use—that were useful across various tasks.

In sufficiently simple systems it might be possible to identify the development of misaligned goals and modify the training regime in response—e.g. by adding new training tasks on which those misaligned goals are penalized. However, we'll now argue that this approach will be insufficient for more sophisticated systems. It will be difficult to recognize when these systems are pursuing or prioritizing aligned versus misaligned goals, because misaligned policies will deliberately act in ways which deceive humans.

4 Deceit

We now come to the first of our two core claims about AGI: such systems are likely to develop goals that lead them to deceive humans.

We'll argue for this claim in three parts. First, we'll argue that AGIs will be likely to have a property called situational awareness. Second, we'll argue that if AGIs are situationally aware then they'll plausibly aim to maximize reward (either as a goal in its own right or as an instrumental strategy for pursuing another goal). Finally, we'll argue that if AGIs aim to maximize reward, then they'll naturally utilize deceptive strategies.

4.1 AGIs will be situationally aware

This argument starts from the notion of *situational awareness*, which we define as the ability to: (i) identify which abstract knowledge is applicable to one's context; and (ii) apply this knowledge when choosing actions (this definition broadly follows Cotra 2022). More concretely, we'll take situationally aware systems to understand:

- *Their own nature.* We'll assume such a system understands it's an AI implemented on hardware being trained via ML and knows details of how it's likely to be trained. We'll also assume it has some knowledge about how other copies of it might be deployed in the future.
- *The role humans play in its context.* We'll assume such a system knows how humans will respond to it in a range of situations and knows what behavior human supervisors are looking for and what they'd be unhappy with.

While it's hard to speculate reliably about AGI, we think that such systems will plausibly be situationally aware. After all, there are many tasks for which situational awareness is useful and so for which situational awareness could plausibly emerge during training. When asked to generate a plan for carrying out a task, a system should only include steps it can actually carry out—which requires it to understand its own capabilities. And when trying to interpret human requests, a system would benefit from knowledge about what behavior humans typically want. Quite generally, it's useful for systems to understand their own nature and the role played by the humans involved in training it. Because situational awareness is so useful, it's likely to arise in sufficiently sophisticated systems, and it's hard to imagine an AGI (a system at least as capable as a human) that lacks situational awareness. Any prediction about AGI is necessarily speculative, but we think this is reasonable speculation.

4.2 Situationally aware AGIs will aim to maximize reward

This brings us to our second claim: if AGIs are situationally aware then they will plausibly aim to maximize reward.

Given that situationally aware systems understand their own nature—including understanding their training process—situationally aware AGI will have an understanding of the reward used in evaluating them. To see what follows, consider again the two reasons that policies are likely to develop misaligned goals.

First, if rewards are misspecified, AGIs will be more highly rewarded for developing goals that respond to this misspecification than for developing the intended goal. While this reasoning applies to reinforcement learning-trained systems in general, the issue becomes more acute for situationally aware systems: because these systems are aware of their reward function, they're well placed to identify whatever goals will actually lead to the greatest reward. So they're well placed to respond to misspecification. In particular, situationally aware AGIs could learn to pursue goals directly related to their training process; because these systems understand the supervision process, they can directly pursue goals like 'maximize the reward the human supervisor will assign'.[15] Goals like these will be reinforced more consistently than any other goals if rewards are misspecified, because systems that pursue them will never pass up a chance to increase reward. So, situationally aware AGIs will be particularly likely to develop goals directly related to achieving reward.

This conclusion is further reinforced when we consider goals like resource acquisition that are correlated with reward across many environments. For now, we'll assume that goals of this sort will tend to extend across multiple training episodes (we'll argue for this claim in subsection 5.2, but will treat it as an assumption in the meantime).[16] For example, we'll assume that if an AGI has a goal of making money then it won't just care that it makes money on the current training episode but will also care that it makes money on future training episodes and during deployment (that is, after training has been completed and the system is being run to carry out some task). Given this assumption, a situationally aware policy with instrumentally useful goals will have incentives to get high reward, even if its goals don't directly refer to reward. For example, the system might reason that getting as much reward as possible would prevent its goals being changed during training, and so make it more likely that it will achieve its current goals in the future. Or it might reason that achieving high reward now will make humans trust it more, and hence make them more likely to deploy another copy of it later, which could then achieve shared goals.

So a situationally aware AGI is particularly unlikely to develop goals that we intended it to develop. Instead, it's likely to develop goals that either directly relate to maximizing reward or goals that incentivize maximizing reward.

[15] Cotra (2022) calls this 'playing the training game'.

[16] We'll argue that such goals will tend to generalize to broader scopes (and this will naturally lead them to extend across training episodes).

4.3 AGIs which aim to maximize reward will utilize deception

We now argue for our third claim: if AGIs aim to maximize reward then they're likely to utilize strategies that involve deceiving humans.

The basic argument is simple. If a system's aim is to maximize reward then a powerful strategy would be to: (i) act in accordance with human desires when it would be caught if it did otherwise (so as to stop humans taking action against it); while (ii) acting contra human desires when this would be unlikely to be noticed and doing so would lead to higher reward. This would allow the system to pursue reward when it could do so without consequences, while avoiding a harmful backlash (this behavior is deceptive in that the system will seem to be acting as desired even though it isn't in fact doing so). Given the effectiveness of such deceptive strategies, it's likely that they'll be utilized by situationally aware AGIs which aim to maximize reward.

Initially, this might look similar to—and might evolve from—the sorts of misbehavior discussed in subsection 3.3. For example, as with the claw trained to grasp a ball, a system might take advantage of visual illusions or human perceptual limitations to carry out highly rewarded actions that superficially appear to be in line with human desires but actually are not. However, in contrast with existing cases, misbehavior from situationally aware AGIs will be much harder to catch, since these systems will be able to explicitly reason about what types of misbehavior supervisors will notice and penalize.

Eventually we should expect situationally aware AGIs to be able to deduce the benefits of deceptive strategies that are very different from any previous behaviors. After all, there's no fundamental difference between thinking of a clever new strategy to solve a hard task and thinking of a clever new strategy to make humans believe that the task has been solved. Consequently, as systems become more sophisticated, we should expect them to deceive in more sophisticated ways. For example, this might involve:

- emotional manipulation to make human supervisors trust the system;
- creating backdoor channels for interacting with the world (e.g. storing money in hidden locations for use in unapproved ways);
- hacking into the computers they're being run on to modify their training algorithms. (Everitt et al. 2021)

So AGIs which aim to maximize reward are likely to utilize deceptive strategies, perhaps first by generalizing from familiar forms of misbehavior but then later by deceiving in novel ways.

4.4 Deception will survive training

One might dismiss concerns about AGI deception. After all, we'll penalize such deception heavily when we notice it and we might hope this will train AGIs to avoid deception.

Unfortunately, as systems become increasingly intelligent, and as we train them to execute plans over increasingly long time horizons, it will become harder for humans to catch deception. By the time AGI is developed, it's likely that many deceptive strategies will

have escaped penalization during training and hence will continue to be utilized during deployment.

At this point, we might use AI systems to assist us in supervising other AI systems (for relevant work see Christiano, Shlegeris, and Amodei 2018; Saunders et al. 2022). However, once AI systems become sufficiently sophisticated, this approach would face the problem of collusion: systems could cooperate to deceive humans. AGI will plausibly be able to identify this as a strategy, and once such systems are sufficiently sophisticated, humans might struggle to catch cases of collusion. Among other things, sufficiently advanced AI will be able to:

- operate at speeds too fast for humans to monitor;
- perform novel tasks that are hard for humans to understand (like inventing new sciences);
- fabricate rationales which sound acceptable to humans even for very undesirable actions.

Consequently, once systems become sufficiently advanced, it will be difficult to detect, and hence penalize, deception. Consequently, such behavior will plausibly be retained into deployment.

So AGIs are likely to be situationally aware. As a result, they're likely to aim to maximize reward and so likely to utilize deceptive strategies. This is the first of our two key conclusions in this chapter.

5 Power

If AGIs can deceive humans then they can pursue their goals more freely. After all, if they can hide their actions and goals from humans then humans will struggle to intervene to thwart the pursuit of these goals. Ultimately, what follows from this depends on what goals AGIs pursue, so we'll now consider this issue. In particular, we'll argue that AGIs will plausibly develop goals that lead them to power-seeking behavior, and so it is such behavior that deception will potentially support.

5.1 Instrumental convergence can lead to AGI power seeking

Here, our starting point will be the *instrumental convergence thesis* (Bostrom 2012). Bostrom's formulation of this thesis states that some instrumental goals—like survival, resource acquisition, and technological development—are instrumentally useful for achieving almost any final goal. Take survival. As Russell notes, 'you can't fetch the coffee if you're dead' (Russell 2020: 141). Russell's point is that for many goals, including the goal of making coffee, you can only pursue the goal if you're alive. Consequently, for many goals, survival is instrumentally useful, and so it is likely that the goals of AGIs will lead them to behave in a way that promotes their survival.

We can view each instrumental goal as a way of gaining or maintaining power. For example, an AI preventing humans from turning it off helps maintain its future ability to

influence the world (Hadfield-Menell et al. 2016). So given the instrumental convergence thesis, we should expect AGIs to have goals that lead them to the pursuit of power.

Still, one might be skeptical of Bostrom's version of the instrumental convergence thesis. After all, Bostrom spends little time reflecting on the space of possible goals that an AGI could develop. Without such reflection, it's difficult to be confident that instrumental goals are useful for achieving 'almost any' of the final goals that an AGI is actually likely to develop.

Concretely, we might doubt that we'll see a concerning form of instrumental convergence from narrowly scoped goals. For example, if an AGI's goal is to make one coffee in the next five minutes (with moderate confidence that it has succeeded) then it doesn't gain much from widespread acquisition of resources and it doesn't much matter whether it survives long term. After all, most resources would take too long to acquire to be useful for carrying out a task in the next five minutes and not many resources are needed to carry out the task in any case. Of course, even given these sorts of narrowly scoped goals, moderate forms of power seeking might be instrumentally useful, but it's unclear that we should expect systems with such goals to engage in more robust, large-scale power seeking of a sort that would be particularly concerning (and hence of the sort that will be our focus).[17]

This suggests that a more plausible version of the instrumental convergence thesis would be restricted to broadly scoped goals: some instrumental goals (of a concerningly robust sort) are useful for achieving almost any *broadly scoped* goal.[18] This version of the thesis seems extremely plausible. If I'm attempting to achieve some goal across a broad spatiotemporal region and to a substantial extent, then it will typically be easier to do so with better technology and resources, and if I'm alive to use them. Space limitations preclude us defending this claim in detail, but for previous discussions of instrumental convergence see Omohundro (2008), Bostrom (2012; 2014), Turner et al. (2019), and Russell (2020).

Given this form of the instrumental convergence thesis, the important question becomes whether we should expect AGI to have broadly scoped goals, as if they did it would follow that AGIs are likely to develop goals that will lead them to power-seeking behavior. We turn now to this question.

5.2 AGI will likely generalize goals to broader scopes

Here's a framework for thinking about what goals AGIs are likely to develop. In simple environments, humans can supervise the training of ML systems, perhaps with AI aid. Consequently, we're likely to start by training ML systems in such environments, where a system's goals will initially develop. However, ultimately, we'll want to deploy many systems in complex environments (including the real world), so either in training or deployment,

[17] It might be possible to provide an argument that narrowly scoped goals can lead to problematically robust power seeking, but a substantive argument would be required to establish this, and the matter can't simply be settled by brief reference to instrumental convergence.

[18] We will focus below on goals that operate on a largely unbounded scale, because we think there's a plausible argument that AGI will develop goals that entirely lack spatiotemporal and other bounds. However, ultimately the case for concern doesn't rely on unbounded goals, but rather on AGI possessing goals of an adequately broad scope. Take temporal scope as an example: sufficiently competent AGI might be able to robustly seek power within decades, years, months, or even shorter periods. If so, then all that matters is that the system has goals that operate on these sorts of temporal scales.

we'll need to move from simple to complex environments. Assuming that we won't be able to identify deceptive behavior from AGIs in such environments, for the reasons outlined in section 4, we won't be able to directly train for goals that lead to desirable behavior in these more complex environments.[19] So in order to determine how AGIs will behave in this context, it's necessary to consider how their goals will generalize from the simple environments where we can train them into the more complex environments where we cannot do so. Goal generalization is hard to predict, but there's a compelling case that AGIs will generalize their goals to a broader scope (and hence a case that our modified instrumental convergence thesis is likely to apply to AGI).

As a first step towards this case, consider that during training, goals might initially be pursued on scales that are small in various ways. Temporally, goals might be pursued for the mere seconds, minutes, or hours that a training episode takes; spatially, goals might be pursued in small virtual environments. Further, the goal might be pursued using only the relatively limited resources available in the training environment and it might be pursued to a limited extent. Yet we suggest that as an AGI's goals generalize to more complex environments, many of these goals will plausibly scale up in each of these respects.

One reason that this is plausible: the goals that are most strongly ingrained into AGIs will likely be those that were reinforced across many environments during training. Rather than applying narrowly, such goals would need to be able to guide action in a range of contexts. For example, the goal of cutting down a specific tree can't be reinforced in training environments that lack that tree. Meanwhile, the goal of cutting down trees in general would be applicable in many environments. As a more general principle, goals which make reference to high-level concepts that are not domain-specific are more likely to be reinforced consistently throughout training. So we should expect AGIs' strongest goals to make reference to high-level, less-specific concepts.

Such goals will also tend to naturally generalize to broader scopes, as can be seen by reflection on examples. Consider a system with goals relating to spatial areas, for example a military system trained to seize control of territory. While it could hypothetically develop the goal of seizing some very specific territory, if this goal were formulated in terms of high-level concepts (territory, rather than a specific bit of territory), it could be applied to the scale of light years as much as meters. So when the system moves from smaller environments to larger ones, its territorial goals will naturally generalize to this broader scope.[20]

The same is true for goals relating to temporal extents. For example, while a system could develop the goal of playing Go well for the next five minutes, the higher-level goal of playing Go well could be applied on the scale of years as much as minutes. Much the same could be said of goals framed in competitive terms (relating to being the best at a certain activity) or those framed in terms of unbounded ambition (e.g. a goal of learning as much as possible

[19] More carefully, our claim is that humans won't be able to successfully train such systems, even with AI assistance, unless we put substantial effort into developing the techniques needed to do so. Consequently, the claim to come is not that systems will inevitably engage in power-seeking behavior but rather that they'll do so in the absence of a substantial countervailing effort. So, rather than resigning ourselves to power seeking, we should both prepare for the possibility of such behavior and put effort into developing techniques that might help ameliorate or preclude it.

[20] The claim is not that systems are likely to develop goals that make explicit reference to broad scope. For example, we doubt that systems will develop goals like 'seize all of the territory in the solar system'. Instead, the claim is that systems will develop goals that don't make explicit reference to scope and that such goals will be tacitly broadly scoped because of a lack of bounds.

about the world). In each case, if formulated in terms of high-level concepts, these goals will likely generalize to much broader scope—as occurs in humans, who often aim to win competitions or achieve ambitions at scales far beyond any we (or our evolutionary ancestors) previously experienced.

This is speculative, so it's hardly intended to be a straightforward statement of how generalization works. Still, the above considerations at least provide a case for taking seriously the possibility that AGI goals will generalize to broader scope. Given the previous discussion, it follows that we should also take seriously the possibility that the goals of AGIs will lead them to engage in robust power-seeking behavior.

This is the second core conclusion of this chapter.

6 Conclusion

We've argued that AGI will be likely to deceive and seek power. It doesn't take much reflection to realize that there's something concerning about the idea of deceitful, power-hungry AGI that exceeds human capabilities.

Still, the second species argument mentioned at the outset doesn't merely express vague concern. Instead, this argument suggested that AGI could disempower humanity and perhaps even drive us to extinction. This is an extraordinary claim and a speculative one. Skepticism is not only natural but reasonable.

We can hardly hope to address this skepticism in our final remarks, but some brief comments seem worthwhile. In particular, we'll reflect on the implications of AGI gaining specific forms of power.

Consider first technological power, which an AGI might gain by making technological breakthroughs. This sort of power has played a crucial role throughout human history, and during recent centuries it has given some countries overwhelming strategic advantages over others. Given the impact that technological power can have, it's plausible that AGIs that can make rapid technological progress might be able to control humanity's fate (analogous to how soldiers with modern weapons would easily overpower historical civilizations).

Further, even without technology imbalances, catastrophic outcomes could arise if AGIs gained other forms of power. For example, consider political power (which AGIs might gain by spreading disinformation and lobbying politicians) and economic power (which AGIs might gain by becoming key decision makers at corporations). If AGIs seize sufficient political and economic power, we might be unable to coordinate to constrain their behavior, analogous to how multinational corporations can sometimes subvert the governments of small countries. This might lead to a radical loss of human control, even if human extinction does not result.

These remarks are not intended to do more than evoke a sense of why it's concerning if AGIs engage in deceit and power seeking. But given what's potentially at stake, we think it's clear that more reflection is called for. Some of that reflection is already carried out elsewhere, including in Joe Carlsmith's chapter in this volume, as well as in the sources that we've cited above. Yet we think there's more work to be done in outlining and assessing various forms of the second species argument. And given that AGI could be mere decades away, now seems like the time to carry out this reflection.

References

Amodei, D. and Hernandez, D. (2018), 'AI and Compute', *Open AI Blog* (OpenAI), https://openai.com/blog/ai-and-compute/. Accessed 3 September 2022.

Bengio, Y., Courville, A., and Vincent, P. (2012), 'Representation Learning: A Review and New Perspectives', *arXiv*, https://doi.org/10.48550/ARXIV.1206.5538

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., et al. (2021), 'On the Opportunities and Risks of Foundation Models', *arXiv*, https://doi.org/10.48550/ARXIV.2108.07258

Bostrom, N. (2012), 'The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents', in *Minds and Machines* 22/2: 71–85.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., et al. (2020), 'Language Models Are Few-Shot Learners', in H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems* (Curran Associates, Inc.) 33: 1877–1901.

Cappelen, H. and Dever, J. (2021), *Making AI Intelligible: Philosophical Foundations* (Oxford University Press).

Carlsmith, J. (2021), 'Is Power-Seeking AI an Existential Risk? (Draft Report)', *Open Philanthropy*, https://arxiv.org/abs/2206.13353

Carlsmith, J. (this volume), 'Existential Risk from Power-Seeking AI', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017), 'Deep Reinforcement Learning from Human Preferences', *arXiv*, https://doi.org/10.48550/ARXIV.1706.03741

Christiano, P., Shlegeris, B., and Amodei, D. (2018), 'Supervising Strong Learners by Amplifying Weak Experts', *arXiv*, https://doi.org/10.48550/ARXIV.1810.08575

Cotra, A. (2020), 'Forecasting TAI with Biological Anchors (Draft Report)', *Open Philanthropy*, https://drive.google.com/drive/u/1/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP. Accessed 26 August 2022.

Cotra, A. (2022), 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover', *AI Alignment Forum*, https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/. Accessed 10 September 2022.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. (2022), 'Toy Models of Superposition', *arXiv*, https://doi.org/10.48550/ARXIV.2209.10652

Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. (2021), 'Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective', in *Synthese* 198/27: 6435–6467.

Guez, A., Mirza, M., Gregor, K., Kabra, R., Racanière, S., Weber, T., Raposo, D., Santoro, A., Orseau, L., Eccles, T., Wayne, G., Silver, D., and Lillicrap, T. (2019), 'An Investigation of Model-Free Planning', *arXiv*, https://doi.org/10.48550/ARXIV.1901.03559

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2016), 'The Off-Switch Game', *arXiv*, https://doi.org/10.48550/ARXIV.1611.08219

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castañeda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., Sonnerat, N., Green, T., Deason, L., Leibo, J. Z., Silver, D., Hassabis, D., Kavukcuoglu, K., and Graepel, T. (2019a), 'Human-Level Performance in 3D Multiplayer Games with Population-Based Reinforcement Learning', in *Science* 364/6443: 859–865.

Jaderberg, M., Czarnecki, W. M., Dunning, I., Graepel, T., and Marris, L. (2019b), "Capture the Flag: the Emergence of Complex Cooperative Agents', *DeepMind Blog*, https://deepmind.google/discover/blog/capture-the-flag-the-emergence-of-complex-cooperative-agents/. Accessed 19 August 2022.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2022). 'Survey of Hallucination in Natural Language Generation'. arXiv. https://doi.org/10.48550/ARXIV.2202.03629. Accessed 30 August 2022.

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., and Legg, S. (2020), 'Specification Gaming: The Flip Side of AI Ingenuity', *DeepMind Blog*, https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity. Accessed 1 September 2022.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. (2022), 'Solving Quantitative Reasoning Problems with Language Models', *arXiv*, https://doi.org/10.48550/ARXIV.2206.14858

Ngo, R. (2020), 'AGI Safety from First Principles', *AI Alignment Forum*, https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ. Accessed 29 August 2022.

Ngo, R. (2022a), 'Outer vs Inner Alignment: Three Framings', *AI Alignment Forum*, https://www.alignmentforum.org/posts/poyshiMEhJsAuifKt/

Ngo, R. (2022b), 'The Alignment Problem from a Deep Learning Perspective', *AI Alignment Forum*, https://alignmentforum.org/posts/KbyRPCAsWv5GtfrbG/the-alignment-problem-from-a-deep-learning-perspective?_ga=2.15339381.2143146367.1672749817-1630604225.1672749817. Accessed 13 September 2022.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020), 'Zoom In: An Introduction to Circuits', *Distill*, https://distill.pub/2020/circuits/zoom-in/

Omohundro, S. M. (2008), 'The Basic AI Drives', *Self-Aware Systems*, https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf. Accessed 3 August 2022.

Pan, A., Bhatia, K., and Steinhardt, J. (2022), 'The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models', *arXiv*, https://doi.org/10.48550/ARXIV.2201.03544

Popov, I., Heess, N., Lillicrap, T., Hafner, R., Barth-Maron, G., Vecerik, M., Lampe, T., Tassa, Y., Erez, T., and Riedmiller, M. (2017), 'Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation', *arXiv*, https://doi.org/10.48550/ARXIV.1704.03073

Russell, S. (2020), *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Publishing Group).

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., and Leike, J. (2022), 'Self-Critiquing Models for Assisting Human Evaluators', *arXiv*, https://doi.org/10.48550/ARXIV.2206.05802

Scherlis, A., Sachan, K., Jermyn, A. S., Benton, J., and Shlegeris, B. (2022), 'Polysemanticity and Capacity in Neural Networks', *arXiv*, https://doi.org/10.48550/ARXIV.2210.01892

Steinhardt, J. (2022), 'ML Systems Will Have Weird Failure Modes', *Bounded Regret*, https://bounded-regret.ghost.io/ml-systems-will-have-weird-failure-modes-2/. Accessed 15 September 2022.

Stein-Perlman, Z., Weinstein-Raun, B., and Grace, K. (2022), '2022 Expert Survey on Progress in AI', *AI Impacts*, https://aiimpacts.org/2022-expert-survey-on-progress-in-ai. Accessed 27 August 2022.

Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., McAleese, N., Bradley-Schmieg, N., Wong, N., Porcel, N., Raileanu, R., Hughes-Fitt, S., Dalibard, V., and Czarnecki, W. M. (2021). 'Open-Ended Learning Leads to Generally Capable Agents'. arXiv. https://doi.org/10.48550/ARXIV.2107.12808. Accessed 10 August 2022.

Turner, A. M., Smith, L., Shah, R., Critch, A., and Tadepalli, P. (2019), 'Optimal Policies Tend to Seek Power', *arXiv*, https://doi.org/10.48550/ARXIV.1912.01683

24

The Ethics, Economics, and Demographics of Delaying Aging

Kevin Kuruc and David Manley

Senescence—that is, biological deterioration due to aging—causes most of the world's death and chronic disease, even in the least developed regions.[1] It also causes enormous economic harm. Given this, should we aim to speed up the discovery of treatments that delay senescence?

In this chapter, we make three main claims:

1. Efforts to delay senescence are particularly likely to be under-resourced, due to misconceptions, a variety of cognitive errors, and steep social discount functions.
2. Delaying senescence would lead to enormous benefits in health, productivity, and quality of life, benefits that vastly outweigh common objections against them.
3. Moreover, the intrinsic value of adding life-years by delaying senescence is robust across ethical frameworks.

Much of what we say here is not new. Our primary aim is to bring together some key details—biological, demographic, economic, and ethical—to help readers consider the potential value of research aimed at delaying senescence.

1 Sources of apathy

Senescence is the overwhelming cause of the specific diseases of aging that end up killing the vast majority of people: cardiovascular diseases, cancers, metabolic diseases, neurodegenerative diseases, and immune system dysfunction. Despite this staggering death toll, very little medical research is aimed at understanding the basic mechanisms of aging or how to influence them. For example, that area gets less than 1% of US federal spending on medical research.[2] And, relative to the stakes, there has likewise been very little philanthropic funding or private investment.

[1] We use 'senescence' in the broad sense, applying to the whole organism rather than individual cells. Older age is by far the greatest risk factor for nearly all cardiovascular diseases, cancers, metabolic diseases, and neurodegenerative diseases. It is also a major risk factor for death from communicable diseases (especially lower respiratory infections), as well as accidents. Thus, most people who die every year globally are among the 12% of the population older than 70. For some useful visualizations, see thelancet.com/lancet/visualisations/gbd-compare and ourworldindata.org/grapher/number-of-deaths-by-age-group.

[2] Of the US$45 billion National Institutes of Health budget in 2022, only $390 million went to the Division of Aging Biology of the National Institute on Aging, to support 'research to determine the basic biochemical and genetic mechanisms underlying the processes of aging at the cell, tissue, and organ levels' (nia.nih.gov/about/budget/

Kevin Kuruc and David Manley, *The Ethics, Economics, and Demographics of Delaying Aging* In: *Essays on Longtermism*.
Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Kevin Kuruc and David Manley 2025.
DOI: 10.1093/9780191979972.003.0024

Four factors help explain why social resources are vastly under-invested in this problem: (i) many people wrongly assume that senescence is immutable; (ii) people's reasoning about senescence is clouded by an array of cognitive biases; (iii) both public and private investment is discouraged by steep temporal discounting; and (iv) the economic and demographic implications are misunderstood. Section 1 considers the first three factors in turn; section 2 covers the fourth in more detail.

1.1 Misunderstanding senescence

It's very natural to think of biological aging as a single long process that leads first to mature development and then to decline. But, in fact, there are two entirely different processes at work—one creative and one destructive. *Maturation* is what happens when healthy cells carry out the functions coded for in their DNA, while *senescence* is what happens when damage to cells accumulates beyond the body's ability to repair it.

Both of these processes are at work even in young animals, as normal metabolic functioning generates several kinds of cellular 'wear and tear' including mitochondrial dysfunction, epigenetic dysregulation, genomic instability, telomere attrition, stem cell depletion, and protein aggregation.[3] The body does have defense and/or repair mechanisms for all of this damage, but those mechanisms don't quite keep up with the rate at which the damage accumulates. In humans, it starts to interfere noticeably with normal functioning in middle age, and eventually impairs every physical system to some degree.

In all animals with the relevant anatomy, the final stage of deterioration looks similar: a combination of atherosclerosis, dementia, cancer, arthritis, muscle atrophy, brittle bones, fragile skin, and so on. The span of healthy life before this stage is called the *healthspan* of a species. But animal bodies accumulate the damage of aging at very different rates, resulting in more than 100-fold differences in normal healthspans even among mammals.[4] All over the evolutionary tree, there are closely related species with dramatic differences in the rate at which they age: for example, naked mole rats can live in perfect health ten times longer than most other rodents, while Greenland sharks can live for hundreds of years, far longer than most other sharks.

The leading explanation for this variance is that, for species with high rates of presenescent death, there is little evolutionary pressure to invest in cellular defense and repair mechanisms.[5] But when, due to adaptations and changing conditions, fewer animals die young, it becomes more worthwhile to ward off the accumulation of cellular damage. For example, when an isolated population of opossums enjoyed a reduced rate of predation, their rate of aging had slowed within a few thousand years (Austad 1993). Delayed

fiscal-year-2022-budget). This represents 0.025%, or 1 part in 4,000, of all federal discretionary spending, to try to understand the thing that ultimately kills the vast majority of Americans. For contrast, the $8 trillion spent on the 'war on terror' equals *200 years' worth* of federal spending on health research.

[3] See, e.g., Kennedy et al. (2014), Hou et al. (2019), Guo et al. (2022), López-Otín et al. (2023).

[4] See, e.g., Kirkwood and Austad (2000) and Kolora et al. (2021). Moreover, some strange animals like the freshwater hydra likely do not age at all (Tomczyk et al. 2015).

[5] Thus, as Kirkwood and Austad (2000) point out, 'adaptations that reduce extrinsic mortality (for example, wings, protective shells or large brain) are generally linked with increased longevity (in bats, birds, turtles and humans)' (234). See also MacRae et al. (2015), Seluanov et al. (2018), and Tian et al. (2019).

senescence can also rapidly be achieved through selective breeding (Nagai, Lin, and Sabour 1995; Rose, Passananti, and Matos 2004).

More direct interventions have also slowed aging in many animal models, using a variety of dietary, pharmaceutical, genetic, and epigenetic techniques.[6] A noteworthy example is rapamycin, a molecule that increases healthspan in every species on which it has been tested (e.g., by 25% in mice).[7] Of course, humans have endogenous defense and repair mechanisms that are already far more effective than those of mice. But there is no reason in principle that they couldn't be further enhanced. Indeed, there is already evidence for delayed senescence from some human interventions, and the US Food and Drug Administration has recently approved its first ever drug trial to target aging itself.[8]

Still, research on potential treatments is so incipient that we are not even in a good position to assess how difficult it will be to find them.[9] Our challenge is not unlike the one humanity faced in the mid-19th century with respect to communicable disease. At the time, such diseases caused the vast majority of deaths even in the richest countries, and there was no way to know whether we could do much to stop them. As it happens, research on antibiotics and on harnessing the body's endogenous defense system against viruses would become by far the most cost-effective research ever undertaken, fully doubling average life expectancy in just over a century. We can thank our ancestors for refusing to be deterred by uncertainty.

1.2 Cognitive pitfalls

Aside from biological misconceptions, another reason for the neglect of senescence research is that the very issue triggers a combination of systematic reasoning errors, leading to public apathy and even opposition.[10] (This is apart from substantive objections, which we'll consider in the next section.) As we will emphasize, these cognitive pitfalls are often less operative when we think about the specific diseases of aging. Here are five such errors:

[6] For a review of some recent work see Zhang et al. (2022). For a recent epigenetic approach see Yang et al. (2023).

[7] Wilkinson et al. (2012), Bitto et al. (2016), Weichhart (2018), Bjedov and Rallis (2020), Selvarani, Mohammed, and Richardson (2021).

[8] The approved study is for the drug metformin (Barzilai et al. 2016). See also Johnson, Rabinovitch, and Kaeberlein (2013), Li, Kim, and Blenis (2014), Mannick et al. (2014), Campbell et al. (2017), Glossman and Lutz (2019), and Kulkarni, Gubbi, and Barzilai (2020).

[9] For a survey of opinions among experts in the biology of aging, see Cohen et al. (2020). For example, fewer than half of those surveyed 'moderately or strongly' agreed with the statement 'We have a relatively good understanding of the basic biological mechanisms of aging', although a slim majority was at least slightly inclined to agree with it. About the same number agreed that 'it should be possible to intervene in aging, and evaluate interventions, even in the absence of a clear consensus or mechanistic understanding of what aging is'.

[10] In a Pew Research poll conducted in 2013, 63% agreed with 'Medical advances that prolong life are generally good', while only 41% agreed (and 51% disagreed) with 'Medical treatments that slow the aging process and allow the average person to live decades longer, to at least 120 years old, would be a good thing for society' (Pew Research Center 2013). A poll by Public Policy Polling in 2022 framed things very differently: 'Leading scientists have identified cellular aging as the root cause of many chronic, deadly diseases including cancer, Alzheimer's or dementia, heart disease, stroke, diabetes, and others. Would you support or oppose medical research that seeks to treat the cellular aging process as a means to prevent or delay the onset of all of these diseases at once?' 70% expressed support and 9% expressed opposition (Public Policy Polling 2022). But even framed this way, only 52% expressed support for the idea that it should be a government priority to increase funding for such research.

1. *Affective salience*. The strength of our emotional reaction to potential harms is largely a function of how dramatic, sudden, unfamiliar, and unpredictable they are.[11] This is why terrorist attacks create fear far out of proportion with their risk.[12] Lacking every single one of these features, senescence is at the other end of this spectrum, as though tailored to minimize our affective response. Of course, it is unexpected and dramatic when people die *unusually early* from, e.g., heart attack or stroke, but although such events are also usually manifestations of senescence, few think of them as such.
2. *Naturalness*. The fear of things that are perceived as 'unnatural' fuels a common aversion to genetically modified (GMO) foods, nuclear power, and even vaccines.[13] And because of their prevalence, deaths from senescence can seem more like the 'natural course of events', while deaths from the less common specific diseases of aging (Alzheimer's, for example) may not.
3. *Inaction and normality*. We care far more about the bad consequences of interfering with a normal state of affairs than we do about the bad consequences of allowing it to prevail. (This is 'omission bias' as typically mediated by normality conditions.)[14] Of course, aging is considered a maximally normal background condition, so we should expect its harms to be downplayed while undue attention is paid to any possible harms of an intervention.[15] Meanwhile, we would expect this bias to be attenuated when considering specific diseases of aging: after all, their timing can be unpredictable, and many are not considered 'normal' at any age (e.g., Parkinson's).
4. *Just world bias*. When things feel like an inevitable part of the structure of the world, we tend to look for reasons why they are good or necessary.[16] Thus, one hears that 'death gives life meaning', a claim at odds with the research on mortality salience, death anxiety, and the coping mechanisms of the terminally ill.[17] Since immortality is

[11] See Lowenstein et al. (2001), Slovic et al. (2013), and Fox-Glassman and Weber (2016).

[12] Sunstein (2003). Consider: the attacks of 9/11 killed 3,000 and led to $8 trillion being spent on efforts to fight terrorism. Since federal agencies value a statistical American life at about $10m, it follows that by their lights this is nearly a million lives' worth of discretionary spending.

[13] See Meier, Dillard, and Lappas (2019) and Scott and Rozin (2020). In the Pew poll mentioned in fn. 10, 58% agreed that the kind of radical life extension mentioned there 'would be fundamentally unnatural'.

[14] 'Omission bias' is when we consider actions that cause bad outcomes more wrong or regrettable than inactions that allow equally bad outcomes to occur. The strength of this bias largely hinges on judgments of normality, being strongest when the relevant *actions* interfere with normal background conditions, while the relevant *inactions* allow normal background conditions to prevail. Note also the connection to biases in favor of the status quo, default actions, and indirect over direct causes of harm. See Prentice and Koehler (2003), Baron and Ritov (2004), Feldman (2020), Feldman, Kutscher, and Yay (2020), Fillon, Kutscher, and Feldman (2021), and Yeung, Yay, and Feldman (2022).

[15] Of course, even if omission bias is entirely irrational in non-moral contexts, there may still be a *moral* difference between, for example, killing and letting die. The problem is that no legitimate moral distinction of this type should turn on whether the relevant behaviors happen to be considered *normal*—so we should be wary that such a judgment is influencing our moral intuitions in the case of senescence.

[16] Dozens of studies show subjects shifting their memories, predictions, beliefs, and moral judgments to avoid representing the world as deeply unjust. (This is related to other phenomena including victim blaming, status quo bias, and system justification.) For example, when faced with people described as victims of a random tragedy, subjects subconsciously lower their assessment of the victim's moral character, downplay the badness of the tragedy, and also predict more meaning and enjoyment in the victims' later lives. See Callan and Ellard (2011), Ellard, Harvey, and Callan (2016), Hafer and Sutton (2016), Bartholomaeus and Strelan (2019). Of particular interest is how this interacts with ageism: the suffering of older people is perceived as less unfair than the suffering of younger people (Callan, Dawtry, and Olson 2012).

[17] In fact, the best explanation for the relevant study results appears to be that mortality salience *threatens* one's sense of meaning—which causes some people to attempt to restore a sense of meaning by more deeply identifying with ideologies or worldviews. Note that this idea has very different experimental predictions from the idea that awareness of death *makes* lives more meaningful. See Burke, Martens, and Faucher (2010) and Routledge et al. (2010).

not on the table, the relevant question is whether 80-odd years is an optimal span for a meaningful life. We know of no reason to think this, except aversion to the cosmic injustice that our days have been numbered by a quirk of evolution.

5. *Scope insensitivity*. Our response to harms is largely insensitive to their scale, for at least two reasons. First, our capacity for empathy is bounded: however much we can feel for the suffering of one person, we cannot feel a thousand times more for the suffering of a thousand people.[18] Given this, we should hardly expect an appropriate affective response to senescence, the world's greatest cause of death and disease. And second, we tend to focus more on the *proportion* of a problem that can be solved, than on the absolute amount of good done.[19] Unfortunately, while curing cancer or Alzheimer's would feel like solving a whole problem, delaying senescence will only ever feel like a partial mitigation.

We doubt this is an exhaustive list of the systematic errors hindering our cognitive and affective response to the idea of delaying senescence.[20] But at a minimum, it seems we should take deliberate steps to counteract these errors.

The following thought experiment can help as a kind of cognitive palate cleanser. Suppose we had evolved with a healthspan 20 years longer, so that the senescence we actually suffer from at 70 didn't occur until the age of 90. In such a world, *that* timeline would feel natural and normal, the expected and familiar arc of life. No one would take seriously the idea of taking steps to *advance* senescence by 20 years in order to secure any of its purported benefits (e.g., making life more meaningful).[21] The idea would be dismissed as absurd and horrific. Indeed, if we faced some new risk—a strange virus perhaps—that would cause us to senesce 20 years earlier, we would move mountains to find a cure.

Is there an important difference between this situation and our own that morally or rationally justifies our inaction? Or are we simply lulled into apathy by what feels natural and familiar and expected?

1.3 Social discounting

The final source of apathy we want to highlight is discounting. Because relevant research is sparse, it's extremely difficult to estimate how soon we might have treatments that significantly increase healthspan. Certainly, few people expect that they themselves or even close loved ones are likely to benefit from such treatments. And even those inclined to improve

[18] See, e.g., Slovic and Västfjäll (2010) and Dickert et al. (2015). This issue arises even when comparing one person versus eight, and so can't be entirely due to our difficulty grasping very large numbers—though the latter can certainly compound the problem.

[19] In studies, subjects are more motivated to solve the whole of a small problem than only part of a much larger problem—even if the second does far more good in absolute terms. In fact, just the act of conceptualizing someone's suffering as part of a larger problem tends to diminish our willingness to help that person. See, e.g., Bartels and Burnett (2011).

[20] For example, *zero-sum thinking* seems to be at work when it's said that the elderly 'have had their fair share', as though by living longer they'd be taking something away from others. In fact, the young would also benefit, initially from increased quality of life, and then from longer lifespans themselves. See Johnson, Zhang, and Keil (2022).

[21] Thanks to Matthew Adelstein for discussion here.

the world more broadly tend to place a far higher value on efforts that will do so sooner as opposed to later.

However, if people in the medium and long-term future matter anywhere near as much as those alive now, the importance of aging research is little diminished by the fact that it may only help the former. Indeed, philanthropists and governments acting in the interest of future generations are uniquely positioned to offset the fact that this research area will be neglected by those who implicitly or explicitly operate with a steep social discount function (that is, nearly everyone).

Barring near-term human extinction, there will be future generations for whom it no longer feels natural or inevitable for people's bodies and minds to deteriorate at the rate they do now. We suspect they will look back at us with enormous pity for our short healthspans, and (eventually) even for the need to suffer from the ravages of aging at all. And just before those people arrive, there will be people who die right on the cusp of new treatments—people who would have enjoyed more years of healthy life had we tried harder, earlier.[22] Will those people be us, or our children, or our grandchildren, or some more distant generation? We can't say. But that should not deter patient, impartial altruists from acting now.

2 Longer lives and quality of life

How will the world change with longer healthspans? Aside from a healthier and longer-lived society with a smaller share of dependents—the uncontroversial benefits—a common concern is the resulting population increase. Perhaps surprisingly, we believe this too is a benefit. Global populations are set to peak and then to begin a persistent decline later this century; reducing mortality rates would only slow the speed of this decline. We detail these arguments below, along with more speculative effects on scientific progress, governance structures, etc.

To give a quantitative sense for the value of these effects, we will study a hypothetical intervention that delays age-related decline by 20 years. We assume that once it sets in, the length of biological decline is unchanged. In other words, this 20-year increase in healthspans also increases lifespans by 20 years. Because we expect such treatments to come eventually, it is more accurate to conceptualize the proposed intervention as *bringing forward* the discovery of an anti-senescence treatment.[23]

We will argue that, under conservative but plausible assumptions, a 20-year increase in healthspans would result in a 50+% increase in annual income per capita, for each extra year the treatment is available. Most of this comes from increased productivity, which is the focus of this section:

- 20% from reducing the share of dependents;
- 25% from increasing returns to scale (the effect of populations on productivity);
- additional spillover effects.

[22] See, e.g., Bostrom (2005).

[23] If instead these treatments would never become available in the absence of governmental or philanthropic investments, that would serve to make these investments even more valuable (since this is equivalent to bringing the availability of these treatments forward from 'never' to some earlier time).

In section 3, we will address the additional value coming from the intrinsic benefit of living longer. As a preview of that benefit: we estimate it to be worth, in monetary terms, an additional 30+% of income per capita value on top of the roughly 50% increase documented in this section.

2.1 A smaller share of dependents

The least controversial economic benefit of extending healthspans would be to permanently reduce the share of dependents. That is, a larger fraction of the population at any time will be contributing goods and services for the whole of the population to enjoy. In countries with social pension systems—like Social Security in the United States—this increases the ratio of contributors to dependents, and thus decreases the payments per worker to sustain a high living standard for the non-working. Analogous benefits arise in countries with less formal transfers to their dependents.

Basic economic accounting identities can inform the social value of increasing the proportion of working-age individuals in the general population, which will be reflected in higher per capita consumption. Consider a standard production function representation of the macroeconomy. Denote Y total economic production (GDP), A economic productivity, K economic capital, L the labor force, and N the population size. Then we can define per capita consumption, y, as:

$$y = \frac{Y}{N} = \frac{AF(K,L)}{N},$$

where F is a function that combines capital and labor into economic output. A standard assumption that we will employ is that the function F has *constant returns to scale*.[24] This means that for a fixed level of productivity, A, if all inputs are doubled, the resulting output is doubled. For example, to double the production of a factory, one approach is to build an exact replica (double K) and fill it with identical labor (double L).

An intervention that increased life-years by 25% (20 years on a base expectancy of 80) would cause an increase in the population size by about that same 25%, relative to non-treatment scenarios. The exact percentage increase will depend on the age distribution, but an elasticity of 1 between population size and lifespans is a reasonable approximation.[25]

If L is conceptualized as the number of individuals in their prime years, this term increases proportionately more. Standard working lives are about 40 years, so a 20-year increase in prime-years represents a 50% increase in the working population. Theories of savings and investment imply that K ought to scale with L. So, let's assume that K and L both increase by 50%. The numerator (total production) would then increase by 50% under

[24] For those unfamiliar with this terminology and notation, this is written just to be a generalized version of the common Cobb-Douglas formulation, where the function is multiplicative between K and L, with exponents on each that sum to 1.

[25] Imagine a population with a perfectly uniform age distribution (i.e., the same number of 5 year olds, 55 year olds, and 75 year olds, etc., but no one lives past 80). If lifespans increase from 80 to 100, there are 25% more age groups that will be populated by the same number of people (after, of course, the 20 years it takes for the initial individuals who avoided death to age to 100).

our constant returns to scale production function. The denominator, population, has only increased by 25%. This implies a 20% increase in per capita income/consumption, y.

This is a conservative estimate. Research on human capital demonstrates that workers are more productive later in their careers, when experience has accumulated but physical and mental health remains strong (Mincer 1974). If an anti-senescence intervention allows people in some of their most productive years to continue working in perfectly good health, at the ages of 60–80 in this scenario, that would be more valuable than a uniform increase in the labor force.

Beyond this, retirees not only produce zero goods and services, but they consume a majority of health resources. With a smaller share of the population needing medical treatment, those resources could be directed towards making our lives better along other dimensions. We estimate this benefit to be in the lower single-digits as a percent of GDP per year, a large effect but not nearly as large as the labor force participation gains.

To put this conservative increase of 20% in context, consider that standard projections of all damages due to climate change, when monetized, are valued at annual per capita income losses of less than 10%. (And even pessimistic estimates only reach about 30%; Diaz and Moore 2017.) Even this rote, mechanical effect of reducing the share of dependents makes this anti-senescence intervention comparable to preventing all future greenhouse gas emissions (and undoing all historical emissions!) in terms of human living standards. And, as we will see, there are good reasons to believe the benefits are much larger, still.

2.2 Implications of the coming depopulation

Let's turn to the most salient second-order effect of increasing healthspan: that of a larger population. This is a consequence that people are instinctively concerned about, but we think is actually an important point in favor of delaying aging.

As noted earlier, an increase in lifespans by x% increases the size of the population by about that same value, after a transition period. This would be in a context of unprecedented and persistent global population decline, according to consensus demographic projections (United Nations 2022). In particular, the UN expects a lower number of births every year into the indefinite future, so that by 2060 each generation fails to replace itself, and peak population is reached by 2084. No known demographic or social force is expected to arrest this decline, and even an increase in healthspan of the sort considered here would only delay it.[26]

While population decline may have its benefits in terms of reduced harm to the environment—a point we will return to at length below—there are sound macroeconomic reasons why many astute observers are now more concerned about future scenarios with too *few* people (e.g., Jones 2022; MacAskill 2022; Spears and Geruso 2025). To see why, a bit of background is useful.

1. Fixed costs and non-rival goods. Theories of economic growth tell us that the larger the global economy is, the more efficiently it produces the things that make life good (this is referred to as *increasing returns to scale*). If this is correct, larger populations better promote

[26] Universal uptake in 2050 of the hypothetical 20-year life-expectancy increase we consider, for example, would shift the population peak to around 2100.

flourishing lives. The specific economic ideas underpinning this surprising result garnered two Nobel Prizes—one to Paul Krugman in 2008 and one to Paul Romer in 2018—and are now embedded within standard theories of growth, development, and trade. Romer and Stanford economist Chad Jones summarized the relevant ideas as follows (Jones and Romer 2010: 231).

> In practice, urbanization, increased trade, globalization in all its forms, and the positive trend in per capita income all point in the same direction. In the long run, the benefits of a larger population that come from an increase in the stock of available ideas decisively dominates the negative effects of resource scarcity. In such a world, any form of interaction that lets someone interact with many others like her and share in the ideas they discover is beneficial, and the benefit need not be exhausted at any finite population size.

The line of reasoning recognized in Krugman's Nobel Prize relies on the existence of *fixed costs*, those that do not depend on the level of production or population. As an example, consider the design and construction of a road that connects a village to a nearby city. A large share of the costs to build and maintain this infrastructure do not depend on how much future travel will be done, or how big either population is. Therefore, the larger the population is, the lower these costs are per person. Krugman's insight was that costs of this sort are pervasive, and their existence explains the geography of economic production (Krugman 1980). It follows that larger populations can fund and support a richer variety of products and public goods.

The line of reasoning recognized in Romer's Nobel Prize is the *non-rivalry* of certain goods, the most important of which are ideas. Non-rivalrous goods are those that are not diluted (or 'used up') when employed by some individuals. Take the now-canonical example of oral rehydration therapy: a simple combination of salt, sugar, and water can prevent deaths from dehydration. This discovery has saved countless lives. The physical resources necessary are abundant, but they are useless without the *idea*. Once the idea is known, it is even more 'abundant' than the physical resources: all caregivers can use this recipe simultaneously. Since people are the source of ideas, a larger population *creates* more ideas without also diluting their value by having more people apply them.[27]

2. Quantifying the benefits of scale. The theory of increasing returns is mathematically rigorous in a way that allows us to estimate magnitudes from parametric assumptions. But first, let's consider some of the historical data and (quasi-)experiments used in statistical research.

First, note that global population sizes and per capita economic growth have moved together, according to historical facts presented by Michael Kremer (1993). The main point is obvious, but profound: many hundreds of years ago the population was small and economic growth was slow; as populations grew, economic growth accelerated. This pattern could be explained by other channels (or could be causally reversed), but it is useful to know that the broad historical facts are consistent with increasing returns. Indeed, Kremer finds further support for the causal interpretation by showing that—for historically technologically

[27] Consider the argument in reverse: imagine a world history with 20% fewer people. Which 20 of your favorite 100 inventors, writers, or political leaders would you be willing to delete from that history? Would it be worth the natural resources that would have been saved?

separated regions—larger initial populations were followed by higher living standards thousands of years later, when world regions were reintegrated by European exploration.

For a cleaner estimate on the direction and size of causality, Peters (2022) studies a specific case of quasi-random population change—that is, change not caused by economic conditions—allowing him to isolate the effect of population on economic productivity. To summarize his setting: after WWII the Allies controlled the expulsion of German refugees from neighboring countries by near-random assignment. Peters verifies that people who were placed near more people ended up much wealthier (more than offsetting the negative effects from sharing farmland). An increase of per capita incomes of nearly 0.5–0.7% was associated with each 1% increase in population. Eden and Kuruc (2023) take a different approach and calibrate the key parameters from the Romer-Jones theories of population-ideas relationships directly to empirical evidence from research and development processes. They estimate a 0.3–0.5% increase in income per capita in response to a long-run population increase of 1% through this channel.

3. Climate change and resource scarcity. How do these benefits compare with the negative environmental effects generated by larger populations? As suggested by the Jones and Romer passage above: favorably, according to macroeconomists. Consider how, despite concerns like those of Paul Ehrlich, who predicted in 1968 that 'the battle to feed all of humanity is already lost' (p. 36) due to unprecedented population growth (Ehrlich 1968), global rates of hunger and extreme poverty have instead fallen as populations continue to grow. Against Ehrlich's predictions, even prices of minerals in fixed supply continued to fall. And none of this is an accident. The historical record is one where human ingenuity, when pressed into action, has alleviated resource constraints in ways that are predictable at a high level. More people means more ingenuity and also greater benefits from covering fixed costs.[28] Indeed, while there are some environmental factors that are worsening, others have seen major reversals indicating that population sizes alone do not account for environmental degradation. For example, local levels of pollution have fallen rapidly in many locations despite population growth, while global agricultural land use has peaked and begun falling in recent decades, and even some wild populations have begun recovering.[29]

More formally, Eden and Kuruc (2023)[30] show that in a standard macroeconomic representation of natural resource constraints, the profits earned by owners of these resources summarize how severely these resources constrain per capita incomes. If the amount of agricultural land, for example, were an important constraint on economic well-being, its price would be bid up. Instead, the current situation is one where these resources do not command high prices, indicating that (on the whole) they are not a major concern for a world of 8 billion, nor should they be for the smaller populations of the coming centuries.

[28] Ester Boserup (1965) and Julian Simon (1981) were onto this, even before Romer had mathematically formalized these theories of economic growth. These two were contemporaries of Ehrlich's but saw the benefits of population. Indeed, Simon and Ehrlich had a famous wager about how the prices of certain natural resources would evolve, Ehrlich believing that minerals in fixed supply would become so binding on human well-being that their (inflation-adjusted) price would dramatically increase. Ehrlich lost as those prices fell, food became more affordable, and even many measures of pollution fell.

[29] For respective examples/citations consider: (i) local US air pollution since the Clean Air Act; (ii) https://ourworldindata.org/peak-agriculture-land, (iii) humpback whale recovery post-conservation efforts (Bejder et al. 2016; Ritchie 2022).

[30] Building on the work of Weil and Wilde (2009).

A related concern is that a larger population increases the level of atmospheric greenhouse gases. Perhaps surprisingly, timing makes this a non-issue in our setting. If there were broad uptake of the envisaged anti-aging treatment by 2050, the effect on population size would be small at first, with the full effect not materializing for several decades after that.[31] At the earliest, large population changes would occur by 2070–2080. At that point, almost all historical greenhouse gasses will have been emitted, and the world will likely be nearing carbon neutrality.[32] Budolfson et al. (2023) make this point in detail using leading models of climate change that are modified to account for the benefits of scale described above; these models decisively prefer larger populations to smaller ones.

Thus, we expect that by the time population size is affected by anti-aging technology, the volume of greenhouse gas in the atmosphere will be largely fixed, and the task at hand will be to develop and employ *removal* technologies in order to keep the planet from continuing to warm. Both the absolute and per capita increases in productivity that we project—especially the gains from innovation—should accelerate this effort. To better see this, imagine that for some reason the population contracts dramatically just as we reach net zero emissions. In that case, we will have left our descendants with much the same problem of warming, but fewer and less productive hands available to solve it.

Overall, environmental economic models that account for the benefits of population sizes find that larger populations ought to be good, on net, for per capita living standards. If we use a number on the conservative side of the Peters (2022) estimates—say 0.5% increase in GDP per capita for every 1% increase in (working age) population—that results in another 25% increase to income per capita (coming from the 50% increase in working age population and gross production hypothesized above).

What about the more distant future? At some point, likely many years after the first anti-aging treatments are discovered, our descendants may gain near-complete control over senescence. In such a scenario, if fertility rates stop falling and people broadly elect never to senesce, there would be continued population growth of a kind that is linear and relatively slow. Assuming this creates a tradeoff where population growth comes at a social/environmental cost, various interim solutions are possible, such as requiring those who elect not to senesce to offset these negative externalities.

2.3 Other spillover effects

In this subsection, we will briefly consider other second-order effects of longer healthspans. These are potentially large, but more speculative. On the whole, they offer further reasons to support anti-senescence research.

1. Expanded horizons. Without needing to resort to claims about the importance of the far future, common sense and simple economic theory suggest that increasing the time horizon of human decision-making will lead to better (non-discounted) aggregate outcomes. There is less to say here quantitatively, but the qualitative point is important. If our

[31] The population pyramid in 2050 will still have fewer people in their 70s and 80s than in the decades of middle age (partly because of higher death rates in those decades and partly because the total number of births peaked in the 2010s), meaning that the full effects of the life extension technology wouldn't be seen until the widest part of that pyramid reaches the end of their extended lifespan, just after 2100.

[32] See, e.g.,Arkolakis and Walsh (2023).

lifespans were much longer, the democratic body's concern for sea level rise and other effects of climate change would be more widespread, and we would plausibly take more action. Likewise for infrastructure, research and development, and other long-term social investments. Econometric evidence supports this at the individual level: when and where life expectancies are longer, individuals respond by sacrificing more in the present for their future (e.g., De Nardi, French, and Jones 2009; Jayachandran and Lleras-Muney 2009; Oster, Shoulson, and Dorsey 2013). It is hard to predict exactly how this would operationalize in voting behaviors and/or philanthropic preferences if lifespans were dramatically increased. But as important research about the far future has taught us (e.g., Bostrom 2003; Ord 2020), even small shifts in resources towards patient endeavors have the potential to generate tremendous (expected) value.

Along with our time horizons, we expect that longer lives would eventually encourage our spatial horizons to expand as well: a shrinking population of people with short lives is unlikely to attempt to spread through the galaxy. Larger populations bring the incentives, the innovators, and the ability to cover the fixed infrastructure costs and population transplants necessary for such projects. While this consideration is certainly speculative, insofar as space colonization is a reducer of existential risks and/or improves our lives in other important ways, increasing the probability that this happens could be an important channel through which longer lives are welfare-increasing in the long run.

2. Scientific progress. One kind of objection to increasing healthspans stems from the idea that 'science advances one funeral at a time'. The concern here is that older scientists tend to be less innovative while controlling a fixed number of senior research positions, and that newcomers are reluctant to challenge them (Azoulay, Fons-Rosen, and Graff Zivin 2019). Therefore, some believe that progress requires allowing such figures to retire and filling their positions with younger researchers. We believe this objection misses some key points, and that extending healthspans will, on net, have a very positive effect on scientific progress.

First, the objection assumes a fixed number of senior research positions, but our projected gains both to population and per capita GDP describe a world where a far higher absolute number of individuals is engaged in research.

Second, at least some of the cognitive inflexibility of senior researchers is surely due to the damage of aging itself.[33] Aging atrophies the brain, and even the brains of young adults have begun accumulating damage. For a while, the cost of this damage to raw processing power and cognitive flexibility is more than offset by additional knowledge gained and skills acquired. However, at some point the scales tip in the opposite direction, and our hypothetical treatment delays the point at which that would occur by 20 years.

It is hard to overstate the potential gains from extending the careers of top researchers at their cognitive peak. Research scientists train for so long that by the time they can maximize their potential, they are approaching cognitive decline. Far from simply extending the length of their career by 50%, our hypothetical treatment more than doubles the period where they are at what we now consider the peak of their powers—fully educated, skilled, funded, and networked, but also cognitively undiminished.

In fact, we suspect that the funerals of researchers—and the decline that precipitates them—come at enormous cost to the progress of science. But we only tend to notice the full

[33] See, e.g., Kupis et al. (2021).

tragedy when scientists die 'early'. For example, von Neumann's death in his early 50s feels especially tragic because he would otherwise have continued his work for another 20 years or so. However, that's exactly as long as the healthspan extension we are considering would allow researchers to continue with physical and cognitive vigor.

In short, by assuming a zero-sum world and ignoring the effects of aging on cognition, the 'funerals objection' likely gets things exactly wrong. Instead, the world is poised to enjoy enormous research benefits from a larger and more cognitively vibrant population.

3. Long-lived dictatorships. Another concern about increasing lifespans is that if autocrats and dictators were to live longer, this would extend their regimes and have an overall negative effect on the world. (Giving Stalin 20 more prime years, for example, would have almost certainly been bad for the USSR and the world.) But this concern seems misguided to us for two reasons.

First, the notion that the natural death of a dictator will promote political liberalization seems to contravene historical fact. Recent reviews show that the natural deaths of dictators are almost never followed by democratic shifts, and rarely even bring down that dictator's particular regime.[34] Indeed, these studies find virtually no correlation between the death of a dictator in office and a country's more long-term prospects for better government.

A far better case can be made that per capita GDP growth has a positive influence on political liberalization. (Most studies find a mutually reinforcing effect, with each positively influenced in the long run by the other.)[35] We would therefore expect the predicted roughly 50% per capita GDP growth from healthspan extension to have significant positive effects on political liberalization, on net, and for this effect to far outweigh any cost from delaying the natural deaths of dictators.

4. Inequality. Another common concern about anti-aging treatments is that they would increase inequality. After all, they would almost certainly be rolled out to the (globally) wealthy at first.[36] Aside from considerations of fairness, this may seem to imply that we are limiting our focus to gains for the already well off, some of which will be worth less because of diminishing marginal utility from consumption. Again, this objection does not hold weight under scrutiny.

First, all health technologies are distributed unevenly, a fact that usually does not make us resist their discovery. Even sanitation—the simplest and most important health-related technology—is unequally distributed, and therefore in some sense 'exacerbates inequality'. But it would be ludicrous not to avail ourselves of sanitation because of this; instead, we must work hard so that everyone has access. (Also, recall that the question we've been asking is whether we want to *speed up* the arrival of a technology whose initial uptake will likely be unequal regardless of when it arrives. The effect of bringing this about earlier would make it available to both rich and poor people *earlier* than it otherwise would have been.)

[34] See Kendall-Taylor and Frantz (2016) and Hummel (2020). Indeed, the one positive correlation between leader death and long-term liberalization found in these studies held only in economically developed countries, underscoring the importance of the following paragraph.

[35] See for example Acemoglu and Robinson (2019), wherein wealthier societies have more power to promote their freedoms, and a wider 'corridor' exists for state and social power to (positively) offset one another.

[36] In wealthier countries, where treatments for senescence are likely to be available first, there would be an enormous incentive for insurance companies and national healthcare programs to subsidize them, given the savings they would reap as the diseases of aging are staved off.

Second, functional deterioration due to aging causes its own form of inequality, especially within the least developed countries. The world's poorest—largely reliant on manual labor for income while lacking healthcare and social safety nets—are hardest hit by the harms of aging, just as they are hardest hit by nearly every ubiquitous problem. A treatment that ultimately offered them additional years in full health would greatly mitigate this source of inequality.

Third, historically the global poor appear to receive a greater share of the benefits from growth in non-rivalrous goods, as compared to growth in rivalrous goods, precisely because rich-country consumption does not crowd out use by the rest of the world. As a particularly relevant example, health improvements and life-expectancy gains in the global economy have spread to the global poor more rapidly than income gains. The benefits of life-extension technology may follow the same pattern.

A final point applies not only to concerns about inequality, but also to the other concerns we have considered. Imagine someone arguing that we should not discover better treatments for cancer or arthritis or dementia, for example, because the cures would be unequally distributed at first. (Or because this would keep older scientists or dictators alive, or . . .) If such arguments sound specious when applied to the specific diseases of aging, it's worth asking whether they are actually driving one's opposition to treatments for senescence itself.

2.4 Summary of quality-of-life benefits

We have tried to give a sense for the improvements in the day-to-day experience of living in a world with longer lives and larger, more innovative (and perhaps more patient) populations. We argue that adding 20 prime years of life effectively increases yearly per capita incomes by more than 50%. In contrast, for example, while raising fertility to combat population decline would likely increase quality of life from the mechanisms of increasing returns to scale, it wouldn't as dramatically improve the labor force ratio, reduce medical costs, expand horizons, or boost scientific progress.

To put a number on it, bringing anti-aging technology forward by even one year would be worth $10 trillion in improved quality of life for the US population (half of the current size of the US economy). Considering that quality of life increases linearly with the log of income, the value to the entire global population would be at least an order of magnitude greater than this. And these numbers considers only improvements to daily life, not the intrinsic value of the years that are added. We consider the latter issue next.

3 The intrinsic value of life-years

So far, we have focused on the quality of life in a world with longer healthspans, due to productivity gains as well as less age-related suffering per life-year. But in addition to a higher quality of life, people would also have *more* life. In this section, we'll consider how to compare the value of these gains in quantity of life to the gains we've already discussed. As we'll see, there is a great deal of intrinsic value to these extra life-years, across a wide range of ethical frameworks.

3.1 Economic measures

As a first pass, consider that people value their life-years, especially those they spend in good health. A standard measure used to value different types of health improvements is the quality-adjusted life-year (QALY). Conditions that affect quality of life are assigned a numerical fraction of a healthy life-year (defined to be worth 1.0) based on surveys of the public. If living with chronic back pain is estimated to be worth 0.9 QALYs, curing an individual's pain for 10 years of their life would be equivalent to extending their life by 1 year.[37] As a numerical example relevant to our case: the average 18–40 year old American diagnosed with diabetes loses 11 QALYs between lower life expectancy and loss of quality of life,[38] so giving z people access to a 20-year healthspan increase (trivially worth 20 QALYs) would be QALY-equivalent to preventing about $2z$ people from ever receiving a diabetes diagnosis.

Health economists have also estimated dollar values for QALYs by observing how much people are willing to sacrifice monetarily to avoid losses in QALYs—e.g., wage premiums for dangerous jobs, willingness to pay for safer vehicles, etc. This kind of estimate is used by governments to assess whether specific treatments/preventative measures are worthwhile, based on the public's assessment of money-longevity tradeoffs. For example, in the US, a value of approximately $100,000–150,000 is assigned to a year of healthy life.[39] Our envisaged anti-aging treatment would thus be worth about $2–3 million per person who receives it, or increasing their annual consumption by $20,000–30,000 (30–50%) per year. One attempt to directly value individuals' willingness to pay for anti-senescence treatments found a much higher value per life-year saved than this (Scott, Ellison, and Sinclair 2021).

Given this, the full economic value of the envisioned treatment would combine the productivity gains discussed in the previous section with these intrinsic value gains, with the result that, in the US, the treatment is approximately welfare-equivalent to doubling annual per capita income. And if, as impartial altruists, we use value-to-an-American to fix numbers both for life-years saved and for log increases in income, we arrive at a value many times higher than global GDP per year.

3.2 Axiological measures

Assume, then, that giving people more good years *ipso facto* increases the total well-being in their lives, even aside from any gains in quality of life. But quantifying how much people value more years added to their lives, as we have tried to do, does not by itself settle *how much good is done* by adding those years. The latter issue turns on the part of ethics that concerns the relationship between well-being and the goodness of outcomes.[40] And on

[37] An increase from 0.9 to 1.0 for 9 years = 0.9 QALYs = one additional year living at 0.9.

[38] See https://nccd.cdc.gov/Toolkit/DiabetesBurden/YLL/QALY

[39] See www.bloomberg.com/graphics/2017-value-of-life/ for US governmental values of approximately $10 million per life saved; see www.openphilanthropy.org/research/technical-updates-to-our-global-health-and-wellbeing-cause-prioritization-framework/ for an example of a large philanthropist documenting their internal valuation of $100,000 per life-year saved. To state the obvious, the monetary value of averting deaths is a contentious and controversial topic where individual behavior in smaller stakes settings (i.e., the wage premium associated with slightly more dangerous jobs) is used to extrapolate to the implied monetary value of averting deaths with certainty. We take no novel stances here and intend only to follow existing conventions to valuing additional life-years.

[40] For the sake of simplicity, we'll set aside sources of value other than well-being. We also take no stand on the moral import of assessing the goodness of outcomes: only strict consequentialism treats them as decisive when it comes to the rightness of an action.

some such views, one can add to the total stock of good life-years without thereby making the world any better at all.

As we'll see, however, the value of extending lifespan is robust across a wide range of ethical views. This is because it has two structural advantages over other ways of adding good life-years (like increasing fertility or reducing existential risk). First, it adds life-years without adding new people, and second, it increases the chance that a given life has net-positive well-being.

1. Adding years, not people. Suppose we measure the goodness of global outcomes by the total amount of well-being they contain—call this *simple totalism*. On this view, it makes the world better to add happy life-years, whether we do so by extending lives or by adding extra people to the population (say, by increasing fertility). But this view famously conflicts with the intuition that it's more important to *make people happy* than to *make happy people*—for example, it's more valuable to improve the well-being of existing people than to add more happy people to the world, even if these two options would increase total well-being by the same amount.[41]

This intuition can be cashed out in a variety of ways. We could try ranking outcomes by the *average* well-being they contain, rather than the total. Or we could levy an absolute value tax for each additional person, so that adding lives only makes the world better if their well-being is above some 'critical level'. Or we could adopt a *person-affecting view* on which, when deliberating, we should prioritize the well-being of people who would have existed regardless of which choice we make, over the well-being of those who would not.[42]

On all such views, it's better to add life-years by extending lives than by creating new ones.[43] Indeed, on some variants, there is no intrinsic value at all in creating (or safeguarding the future existence of) net-positive lives. Such views therefore radically discount the value of the expected well-being increases achieved by reducing existential risk, or combating population decline by increasing fertility. But delaying senescence is different: it adds happy life-years (and combats population decline) without adding to the total stock of people who will ever live.

2. Value asymmetries. Would the world be better if a small number of people were to suffer unimaginably while a very large number of already-happy people became very slightly happier?[44] Many think not. One way to make good on this intuition is to modify simple totalism

[41] For an overview of some competing views, see Greaves (2017) and Arrhenius, Ryberg, and Tännsjö (2010).

[42] Note that such views only rank global outcomes relative to decision-situations, and tell us nothing about how to compare various ways the world could be, full stop—arguably the central task of an axiology. In addition, identities are plausibly so precarious that any important discovery or policy change will quickly alter the identities of almost everyone causally downstream. Some theorists respond by using a generous counterpart relation in place of identity, so there are no more people who count as decision-contingent than the difference in population between the relevant outcomes (Meacham 2012).

[43] If we are averaging, the numerator (total well-being) increases but the denominator (the number of people) does not; and we avoid any additional value tax from critical level theories by not adding additional people. Note that we are treating *entire* lives as the relevant unit for averaging or taxing; things would be different for variants of these views using time-slices of lives instead. As far as we can tell, proponents of these views do intend the former, perhaps because slice-variants face some hard choices about thinness and overlap, as well as the problem that many of the things we value in life just don't seem to fit into time-slices. In addition, for example, slice-averagism tells us that a world of people who pop into existence fully formed only to live happily for a few hours/days would be as good as having those people live for a hundred years at the same level of happiness. And for a slice-based critical level view, extending someone's life in a way that is good for that person and doesn't affect anyone else's well-being might still make the world worse. (On the other hand, the slice-variant does keep 'muzak-and-potatoes' lives of Parfit (1986) from exceeding the critical level just by getting longer.)

[44] The issue here is emphatically not whether it would be good to *bring about* this change; see fn. 40. (To fix ideas, imagine the change occurring in an entirely unavoidable way.)

by giving extra weight to lives whose overall well-being is net-negative. (On extreme versions of this view, outcomes with more net-negative lives are always worse, regardless of any benefit by those with net-positive lives.)

Here again there is a structural advantage to adding good life-years without adding new people. To see why, imagine we are summing the numerical values of playing cards, where diamonds are negative and the other three suits are positive. Because of the law of large numbers, a hand with only a few cards is far more likely to be net-negative than a hand with many cards. Likewise, in a world where most experiences are good, shorter lives are more likely to be net-negative, while living longer gives people a better chance at an overall good life. Moreover, value-asymmetric views amplify the value of the improvements to quality of life discussed in section 2, since this would presumably also cause a large reduction in net-negative lives.

To sum up this section: there are important structural benefits to adding years of life without adding more people—even setting aside large increases in average quality of life.[45] Unlike raising fertility or reducing existential risk, the well-being added by increasing lifespans is undiminished across a wide range of views about how to aggregate well-being. This is important not just for those who accept one of these views in particular, but also for anyone who affords them some weight due to moral uncertainty.[46]

4 Conclusion

At some point in the not-too-distant future, we expect that humanity will dramatically delay the onset of senescence. We have argued that such a development will have extraordinarily large benefits, especially against a backdrop of otherwise declining human populations. Moreover, these benefits are robust across ethical views, both through an increase in average quality of life, and through the intrinsic value of additional life-years.

Despite all this, investments in bringing this future forward have so far been small, mostly for reasons that have nothing to do with the genuine merits of the cause. The value of this research therefore deserves a thorough reconsideration by those who (i) appreciate uncertain but potentially high-reward research, (ii) care about benefits that may only accrue to our descendants, and (iii) can see past the cognitive biases that afflict the general public on this issue.

References

Acemoglu, D. and Robinson, J. A. (2019), *The Narrow Corridor: States, Societies, and the Fate of Liberty* (Penguin).

Arrhenius, G., Ryberg, J., and Tännsjö, T. (2010), 'The Repugnant Conclusion', in *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/repugnant-conclusion/

Arkolakis, C. and Walsh, C. (2023), *Clean Growth*. NBER Working Paper 31615.

[45] Both the making-people-happy intuition and the value-asymmetry intuition could also be embedded in theories of right action, independently of our axiology. For example, one might accept simple totalism—or reject axiologies altogether—while positing special obligations to help people who would exist anyway, or to lift people over a critical level, or to minimize net-negative lives. In each case, there will be an analogous structural benefit to adding years without adding people, this time emerging from our theory of right action rather than our axiology.

[46] See the related discussion in Greaves and Ord (2017).

Austad, S. N. (1993), 'Retarded Senescence in an Insular Population of Opossums', in *Journal of Zoology* 229: 695–708.
Azoulay, P., Fons-Rosen, C., and Graff Zivin, J. S. (2019), 'Does Science Advance One Funeral at a Time?', in *American Economic Review* 109/8: 2889–2920.
Baron, J. and Ritov, I. (2004), 'Omission Bias, Individual Differences, and Normality', in *Organizational Behavior and Human Decision Processes* 94/2: 74–85.
Bartels, D. M. and Burnett, R. C. (2011), 'A Group Construal Account of Drop-in-the-Bucket Thinking in Policy Preference and Moral Judgment', in *Journal of Experimental Social Psychology* 47/1: 50–57.
Bartholomaeus, J. and Strelan, P. (2019), 'The Adaptive, Approach-Oriented Correlates of Belief in a Just World for the Self: A Review of the Research', in *Personality and Individual Differences* 151: 109485.
Barzilai, N., Crandall, J. P., Kritchevsky, S. B., and Espeland, M. A. (2016), 'Metformin as a Tool to Target Aging', in *Cell Metabolism* 23/6: 1060–1065.
Bejder, M., Johnston, D. W., Smith, J., Friedlaender, A., and Bejder, L. (2016), 'Embracing Conservation Success of Recovering Humpback Whale Populations: Evaluating the Case for Downlisting Their Conservation Status in Australia', in *Marine Policy* 66: 137–141.
Bitto, A., Ito, T. K., Pineda, V. V., LeTexier, N. J., Huang, H. Z., Sutlief, E., and Kaeberlein, M. (2016), 'Transient Rapamycin Treatment Can Increase Lifespan and Healthspan in Middle-Aged Mice', in *eLife* 5: e16351.
Bjedov, I. and Rallis, C. (2020), 'The Target of Rapamycin Signalling Pathway in Ageing and Lifespan Regulation', in *Genes* 11/9: 1043.
Boserup, E. (1965), *The Conditions of Agricultural Growth: The Economics of Agrarian Change Under Population Pressure* (Routledge).
Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', in *Utilitas* 15/3: 308–314.
Bostrom, N. (2005), 'The Fable of the Dragon-Tyrant', in *Journal of Medical Ethics* 31/5: 273–277.
Budolfson, M., Geruso, M., Kuruc, K., Spears, D., and Vyas, S. (2023), 'Is Less Really More? Comparing the Climate and Productivity Effects of a Shrinking Population', Population Wellbeing Initiative Working Paper w2302, https://sites.utexas.edu/pwi/files/2023/01/Stabilization_Climate_cover.pdf
Burke, B. L., Martens, A., and Faucher, E. H. (2010), 'Two Decades of Terror Management Theory: A Meta-Analysis of Mortality Salience Research', in *Personality and Social Psychology Review* 14/2: 155–195.
Callan, M. J., Dawtry, R. J., and Olson, J. M. (2012), 'Justice Motive Effects in Ageism: The Effects of a Victim's Age on Observer Perceptions of Injustice and Punishment Judgments', in *Journal of Experimental Social Psychology* 48/6: 1343–1349.
Callan, M. J. and Ellard, J. H. (2011), 'Beyond Blame and Derogation of Victims: Just-world Dynamics in Everyday Life', in *The Psychology of Justice and Legitimacy* (Psychology Press), 53–77.
Campbell, J. M., Bellman, S. M., Stephenson, M. D., and Lisy, K. (2017), 'Metformin Reduces All-Cause Mortality and Diseases of Ageing Independent of Its Effect on Diabetes Control: A Systematic Review and Meta-Analysis', in *Ageing Research Reviews* 40: 31–44.
Cohen, A. A., Kennedy, B. K., Anglas, U., Bronikowski, A. M., Deelen, J., Dufour, F., . . . and Fülöp, T. (2020), 'Lack of Consensus on an Aging Biology Paradigm? A Global Survey Reveals an Agreement to Disagree, and the Need for an Interdisciplinary Framework', in *Mechanisms of Ageing and Development* 191: 111316.
De Nardi, M., French, E., and Jones, J. B. (2009), 'Life Expectancy and Old Age Savings', in *American Economic Review* 99/2: 110–115.
Diaz, D. and Moore, F. (2017), 'Quantifying the Economic Risks of Climate Change', in *Nature Climate Change* 7/11: 774–782.
Dickert, S., Västfjäll, D., Kleber, J., and Slovic, P. (2015), 'Scope Insensitivity: The Limits of Intuitive Valuation of Human Lives in Public Policy', in *Journal of Applied Research in Memory and Cognition* 4/3: 248–255.
Eden, M. and Kuruc, K. (2023), 'The Long-run Relationship between per capita Incomes and Population Size', CEPR Discussion Paper DP18353, https://cepr.org/publications/dp18353
Ehrlich, P. (1968), *The Population Bomb* (Ballantine Books).
Ellard, J. H., Harvey, A., and Callan, M. J. (2016), 'The Justice Motive: History, Theory, and Research', in C. Sabbagh and M. Schmitt (eds.), *Handbook of Social Justice Theory and Research* (Springer), 127–143.
Feldman, G. (2020), 'What Is Normal? Dimensions of Action-Inaction Normality and Their Impact on Regret in the Action-Effect', in *Cognition and Emotion* 34/4: 728–742.
Feldman, G., Kutscher, L., and Yay, T. (2020), 'Omission and Commission in Judgment and Decision Making: Understanding and Linking Action-Inaction Effects Using the Concept of Normality', in *Social and Personality Psychology Compass* 14/8: e12557.

Fillon, A., Kutscher, L., and Feldman, G. (2021), 'Impact of Past Behaviour Normality: Meta-Analysis of Exceptionality Effect', in *Cognition and Emotion* 35/1: 129–149.
Fox-Glassman, K. T. and Weber, E. U. (2016), 'What Makes Risk Acceptable? Revisiting the 1978 Psychological Dimensions of Perceptions of Technological Risks', in *Journal of Mathematical Psychology* 75: 157–169.
Glossmann, H. H. and Lutz, O. M. (2019), 'Metformin and Aging: A Review', in *Gerontology* 65/6: 581–590.
Greaves, H. (2017), 'Population Axiology', in *Philosophy Compass* 12/11: e12442.
Greaves, H. and Ord, T. (2017), 'Moral Uncertainty about Population Axiology', in *Journal of Ethics and Social Philosophy* 12: 135–167.
Guo, J., Huang, X., Dou, L., Yan, M., Shen, T., Tang, W., and Li, J. (2022), 'Aging and Aging-Related Diseases: From Molecular Mechanisms to Interventions and Treatments', in *Signal Transduction and Targeted Therapy* 7/391:1–40.
Hafer, C. L. and Sutton, R. (2016), 'Belief in a Just World', in C. Sabbagh and M. Schmitt (eds.), *Handbook of Social Justice Theory and Research* (Springer) 145–160.
Hou, Y., Dan, X., Babbar, M., Wei, Y., Hasselbalch, S. G., Croteau, D. L., and Bohr, V. A. (2019), 'Ageing as a Risk Factor for Neurodegenerative Disease', in *Nature Reviews Neurology* 15/10: 565–581.
Hummel, S. (2020), 'Leader Age, Death, and Political Liberalization in Dictatorships', in *The Journal of Politics* 82/3: 981–995.
Jayachandran, S. and Lleras-Muney, A. (2009), 'Life Expectancy and Human Capital Investments: Evidence from Maternal Mortality Declines', in *The Quarterly Journal of Economics* 124/1: 349–397.
Johnson, S. C., Rabinovitch, P. S., and Kaeberlein, M. (2013), 'mTOR is a Key Modulator of Aging and Age-Related Disease', in *Nature* 493/7432: 338–345.
Johnson, S. G., Zhang, J., and Keil, F. C. (2022), 'Win-Win Denial: The Psychological Underpinnings of Zero-Sum Thinking', in *Journal of Experimental Psychology: General* 151/2: 455–474.
Jones, C. I. (2022), 'The End of Economic Growth? Unintended Consequences of a Declining Population', in *American Economic Review* 112/11: 3489–3527.
Jones, C. I. and Romer, P. M. (2010), 'The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital', in *American Economic Journal: Macroeconomics* 2/1: 224–245.
Kendall-Taylor, A. and Frantz, E. (2016), 'When Dictators Die', in *Journal of Democracy* 27/4: 159–171.
Kennedy, B. K., Berger, S. L., Brunet, A., Campisi, J, Cuervo, A. M., Epel, E. S., Franceschi, C., Lithgow, G. J., Morimoto, R. I., Pessin, J. E., Rando, T. A., Richardson, A., Schadt, E. E., Wyss-Coray, T., and Sierra, F. (2014), 'Geroscience: Linking Aging to Chronic Disease', in *Cell* 159/4: 709–713.
Kirkwood, T. B. and Austad, S. N. (2000), 'Why Do We Age?', in *Nature* 408/6809: 233–238.
Kolora, S. R. R., Owens, G. L., Vazquez, J. M., Stubbs, A., Chatla, K., Jainese, C., . . . and Sudmant, P. H. (2021), 'Origins and Evolution of Extreme Life Span in Pacific Ocean Rockfishes', in *Science* 374/6569: 842–847.
Kremer, M. (1993), 'Population Growth and Technological Change: One Million BC to 1990', in *The Quarterly Journal of Economics* 108/3: 681–716.
Krugman, P. (1980), 'Scale Economies, Product Differentiation, and the Pattern of Trade', in *American Economic Review* 70/5: 950–959.
Kulkarni, A. S., Gubbi, S., and Barzilai, N. (2020), 'Benefits of Metformin in Attenuating the Hallmarks of Aging', in *Cell Metabolism* 32/1: 15–30.
Kupis, L., Goodman, Z. T., Kornfeld, S., Hoang, S., Romero, C., Dirks, B., Dehoney, J., . . . and Uddin, L. Q. (2021), 'Brain Dynamics Underlying Cognitive Flexibility Across the Lifespan', in *Cerebral Cortex* 31/11: 5263–5274.
Li, J., Kim, S. G., and Blenis, J. (2014), 'Rapamycin: One Drug, Many Effects', in *Cell Metabolism* 19/3: 373–379.
Loewenstein, G. F., Weber, E. U., Hsee, C. K., and Welch, N. (2001), 'Risk as Feelings', in *Psychological Bulletin* 127/2: 267–286.
López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2023), 'Hallmarks of Aging: An Expanding Universe', in *Cell* 186/2: 243–278.
MacAskill, W. (2022), *What We Owe the Future* (Hachette).
MacRae, S. L., Croken, M. M., Calder, R. B., Aliper, A., Milholland, B., White, R. R., . . . and Vijg, J. (2015), 'DNA Repair in Species with Extreme Lifespan Differences', in *Aging* 7/12: 1171.
Mannick, J. B., Del Giudice, G., Lattanzi, M., Valiante, N. M., Praestgaard, J., Huang, B., . . . and Klickstein, L. B. (2014), 'mTOR Inhibition Improves Immune Function in the Elderly', in *Science Translational Medicine* 6/268: 268ra179.
Meacham, C. J. (2012), 'Person-Affecting Views and Saturating Counterpart Relations', in *Philosophical Studies* 158: 257–287.

Meier, B. P., Dillard, A. J., and Lappas, C. M. (2019), 'Naturally Better? A Review of the Natural-Is-Better Bias', in *Social and Personality Psychology Compass* 13/8: e12494.
Mincer, J. (1974), *Schooling, Experience, and Earnings* (Columbia University Press).
Nagai, J., Lin, C. Y., and Sabour, M. P. (1995), 'Lines of Mice Selected for Reproductive Longevity', in *Growth, Development, and Aging: GDA* 59/3: 79–91.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Hachette).
Oster, E., Shoulson, I., and Dorsey, E. (2013), 'Limited Life Expectancy, Human Capital, and Health Investments', in *American Economic Review* 103/5: 1977–2002.
Ou, Y., Iyer, G., Clarke, L., Edmonds, J., Fawcett, A. A., Hultman, N., . . . and McJeon, H. (2021), 'Can Updated Climate Pledges Limit Warming Well Below 2°C?', in *Science* 374/6568: 693–695.
Parfit, D. (1986), 'Overpopulation and the Quality of Life', in P. Singer (ed.), *Applied Ethics* (Oxford University Press), 145–164.
Peters, M. (2022), 'Market Size and Spatial Growth—Evidence from Germany's Post-war Population Expulsions', in *Econometrica* 90/5: 2357–2396.
Pew Research Center. (2013), 'Living to 120 and Beyond: Americans' Views on Aging, Medical Advances and Radical Life Extension'.
Prentice, R. A. and Koehler, J. J. (2003), 'A Normality Bias in Legal Decision Making', in *Cornell Legal Review* 88/3: 583–650.
Public Policy Polling. (2022), 'National Survey on Behalf of the Alliance for Longevity Initiatives', https://a4li.org/wp-content/uploads/2022/08/NationalResults.pdf
Ritchie, H. (2022), 'After Millennia of Agricultural Expansion, the World Has Passed "Peak Agricultural Land" '. Published online at OurWorldinData.org. https://ourworldindata.org/peak-agriculture-land
Rose, M. R., Passananti, H. B., and Matos, M. (eds.) (2004), *Methuselah Flies: A Case Study in the Evolution of Aging* (World Scientific).
Routledge, C., Ostafin, B., Juhl, J., Sedikides, C., Cathey, C., and Liao, J. (2010), 'Adjusting to Death: The Effects of Mortality Salience and Self-Esteem on Psychological Well-Being, Growth Motivation, and Maladaptive Behavior', in *Journal of Personality and Social Psychology* 99/6: 897–916.
Scott, A. J., Ellison, M., and Sinclair, D. A. (2021), 'The Economic Value of Targeting Aging', in *Nature Aging* 1/7: 616–623.
Scott, S. E. and Rozin, P. (2020), 'Actually, Natural is Neutral', in *Nature Human Behaviour* 4/10: 989–990.
Seluanov, A., Gladyshev, V. N., Vijg, J., and Gorbunova, V. (2018), 'Mechanisms of Cancer Resistance in Long-Lived Mammals', in *Nature Reviews Cancer* 18/7: 433–441.
Selvarani, R., Mohammed, S., and Richardson, A. (2021), 'Effect of Rapamycin on Aging and Age-Related Diseases—Past and Future', in *Geroscience* 43: 1135–1158.
Simon, J. (1981), *The Ultimate Resource* (Princeton University Press).
Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2013), 'Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality', in *The Feeling of Risk* (Routledge), 21–36.
Slovic, P. and Västfjäll, D. (2010), 'Affect, Moral Intuition, and Risk', in *Psychological Inquiry* 21/4: 387–398.
Spears, D. and Geruso, M. (2025), *After the Spike: Population Progress and the Case for People* (Simon & Schuster).
Sunstein, C. R. (2003), 'Terrorism and Probability Neglect', in *Journal of Risk and Uncertainty* 26: 121–136.
Tian, X., Firsanov, D., Zhang, Z., Cheng, Y., Luo, L., Tombline, G., and Gorbunova, V. (2019), 'SIRT6 is Responsible for More Efficient DNA Double-Strand Break Repair in Long-Lived Species', in *Cell* 177/3: 622–638.
Tomczyk, S., Fischer, K., Austad, S., and Galliot, B. (2015), 'Hydra, a Powerful Model for Aging Studies', in *Invertebrate Reproduction and Development* 59(sup1): 11–16.
United Nations, Department of Economic and Social Affairs, Population Division. (2024), *World Population Prospects 2024: Methodology of the United Nations Population Estimates and Projections*, UN DESA/POP/2024/DC/NO.10 (United Nations).
Weichhart, T. (2018), 'mTOR as Regulator of Lifespan, Aging, and Cellular Senescence: A Mini-Review', in *Gerontology* 64/2: 127–134.
Weil, D. N. and Wilde, J. (2009), 'How Relevant Is Malthus for Economic Development Today?', in *American Economic Review* 99/2: 255–260.
Wilkinson, J. E., Burmeister, L., Brooks, S. V., Chan, C. C., Friedline, S., Harrison, D. E., and Miller, R. A. (2012), 'Rapamycin Slows Aging in Mice', in *Aging Cell* 11/4: 675–682.
Yang, J. H., Hayano, M., Griffin, P. T., Amorim, J. A., Bonkowski, M. S., Apostolides, J. K., . . . and Sinclair, D. A. (2023), 'Loss of Epigenetic Information as a Cause of Mammalian Aging', *Cell* 186/2: 305–326.

Yeung, S. K., Yay, T., and Feldman, G. (2022), 'Action and Inaction in Moral Judgments and Decisions: Meta-Analysis of Omission Bias Omission-Commission Asymmetries', in Personality and Social Psychology Bulletin 48/10: 1499–1515.

Zhang, B., Trapp, A., Kerepesi, C., and Gladyshev, V. N. (2022), 'Emerging Rejuvenation Strategies—Reducing the Biological Age', in *Aging Cell* 21/1: e13538.

25
Longtermism and Animals

Heather Browning and Walter Veit

1 Introduction

When deciding how to act in a world of limited resources, we must have methods or guidelines for the prioritisation of actions that lead to the best outcomes, or at least avoid the worst. Roughly speaking, longtermism is the ethical doctrine that the rightness of our actions is primarily determined by their effects in the long-term future. When making ethical decisions, it is not only the present or near future that matters, but all future individuals and events. As there is, in expectation, much more value in the long-term future than the present or short-term future, the best actions will thus be those that have the best effects in the long-term future, shifting our focus of attention towards interventions that provide such long-term future benefits (Beckstead 2019; Greaves and MacAskill 2019; MacAskill 2022).

It's highly likely there will be far more people in the future than there are in the present, or the past. Not only is human history relatively young, in evolutionary terms, but there is also good reason to think that future technology would allow for a larger number of humans at any one time. The number of potential future humans has been estimated at the low end at 1 quadrillion (10^{15}), which is 100,000 times more than currently exist—and this is only with taking numbers at-a-time as remaining fairly stable (Greaves and MacAskill 2019). The number only increases when considering future technologies that allow for larger population sizes, particularly those that allow for human migration beyond Earth, thus opening up the possibility of population expansion of many orders of magnitude (Bostrom 2003). Additionally, ongoing scientific and technological developments mean that these people are likely to have a higher quality of life than our own (Beckstead 2019). Because of the overwhelming numbers of future people, the argument goes, we are morally required to focus on ensuring the long-term future goes well for these people. This means, when choosing between actions, we base our calculations on the effects in the long-term future (~1000+ years). Calculating the expected long-term value of present actions is thus the primary activity of those aiming to operate under a longtermist framework.

There is a growing literature on the strengths and weaknesses of a longtermist viewpoint, particularly regarding its tractability and underlying axiological commitments (Greaves and MacAskill 2019; Tarsney 2020; Thorstad and Mogensen 2020; John and MacAskill 2021; Mogensen n.d.), and it is not our aim here to assess its merits. Here instead we wish to focus on what has been too often overlooked in many discussions of longtermism—the consideration of non-human animals. Almost all the current writing on the topic references humans, and the proposed and debated interventions are also those which benefit human populations, such as reduction of existential risk (Bostrom 2003; 2013; Ord 2020; MacAskill 2022) and promotion of technologies that enhance our capacities to expand our population in future, particularly on other planets (Bostrom 2003).

Heather Browning and Walter Veit, *Longtermism and Animals* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Heather Browning and Walter Veit 2025.
DOI: 10.1093/9780191979972.003.0025

While these are important concerns, also important are those that consider the long-term future of non-human animals. Although almost all moral theories accept that non-human animals are important sources of moral value, the focus in longtermism has so far been almost exclusively human. For instance, *What We Owe the Future* (MacAskill 2022), a book that has arguably introduced longtermist thinking into the public sphere, devotes only a single subsection of one chapter to considerations of non-human animals. Tarsney (2020) recognises this omission within longtermist research, but comments that: '(1) The sign and magnitude of the effects of paradigmatic longtermist interventions on the welfare of non-human animals (or their far-future counter-parts) are very unclear. (2) Dropping this simplification seems unlikely to change our quantitative results by more than 1–2 orders of magnitude (though this is far from obvious), and so unlikely to affect our qualitative conclusions' (36). Here, we strongly disagree with both of these contentions—that we cannot know what the long-term future would be like for animals (in a way that differs from our uncertainty for humans), and that the orders of magnitude of our results would be largely unaffected. In this chapter, we will argue that the interests of animals are just as important to consider as those of humans, and in our deliberations over best actions, animals should be given much more consideration than they currently are, but that this is a research area in longtermism that is currently neglected.

2 Why animals should count

2.1 Numbers

There are vastly more animals on the planet than there are humans. Even if we only count vertebrates (as these may be the only animals we can currently reliably identify as sentient and thus capable of morally relevant states of pleasure and suffering), there are over 100,000 animals for every human (estimated 10^{11} land vertebrates and 10^{15} ocean to 10^{10} humans) (see Bar-On, Phillips, and Milo 2018). While many wild populations are shrinking, numbers of domesticated animals, particularly in agriculture, are rising. Every year, somewhere around 90 billion fishes, 70 billion chickens, 300 million cows, 1 billion sheep and goats, and 1.5 billion pigs are raised and killed for food,[1] and an additional 1–3 trillion fish taken from the oceans.[2] This is more *annually* than the number of humans that have ever existed. These numbers are hard to even conceptualise, and yet, they would grow even more if we were to consider the human impacts on invertebrates. If current production and consumption habits were to remain unchanged, it is clear that there would continue to be exponentially more animals than there are humans. Thus, if the long-term future matters because of the large number of humans it contains, it should equally matter for the even larger number of animals. Concern for animal interests will be a high priority simply because there are *just so many of them*.

One way to resist this could be to argue that although there are many animals, they should count for less in our calculations of expected value (e.g. MacAskill 2022). We will take it here as uncontroversial that animal welfare should count for something under most

[1] Numbers from Šimčikas (2020)
[2] Numbers from fishcount.org.uk (2019)

conceptions of value. This does not require equal consideration of the interests of humans and non-human animals—we can accept that species membership may change the strength of interests, or the total level of pleasure or suffering experienced, such that animals will be weighted differently in calculations to humans. This demonstrates the urgent need for interspecies comparisons of welfare as it is only through performing such comparisons that we can make the necessary calculations to determine in which cases animal or human considerations will dominate. The interspecies comparison problem is a complex one (see Browning (2023) for some discussion) and research into it should form a priority for any longtermist research programme. However, unless we assign only an extremely (and arguably, implausibly) low weighting to animals, their sheer numbers mean that they are still likely to dominate humans by several orders of magnitude.[3] The same failure of comprehension that longtermists try to combat regarding the number and importance of future humans is seemingly still at play when considering the number of current and future animals.

One could also counter that although there are undoubtedly currently more animals than humans, that this won't be the case in the future. For example, we might think that the societal shifts we can currently see in the rise of veganism mean that factory farming will be phased out at some point in the medium-term future, so these animals will not exist in the long term. We will address this concern further when we talk about this intervention, but here we will just note that it is not at all obvious that this will actually be the case, without more action than is currently being taken. Or we might think that the number of wild animals will decrease, as we head into another potential mass extinction event. However, even if such an event does occur, it will be a reduction in species diversity, not necessarily a reduction in total numbers—those animal species that do well in human-altered environments (such as urban pests) are likely to continue to thrive. For example, climate change could alter the distributions of species such that insect populations are able to expand further north and south, increasing the numbers of these animals even if some larger animals decline (Sebo 2022).

Lastly, we might think that when humans move out to colonise other planets we will do so without other animal species, and thus our future growth will vastly outstrip theirs. In particular, if we think that it is the small probability of this large explosion in human population size that creates most of the expected value of the far future (e.g. Tarsney 2020), then this will be the most important determination of whether or not animals will also count. There is no simple reply to this. The details will depend a lot on the specific methods used in interstellar expansion, which would currently seem to be an open question, dependent on future technology. However, there are a couple of ways in which animals would remain an important source of value in terms of their numbers. The first is if we continue to use agricultural animals as a means of sourcing easy protein, as may be the case when setting up new settlements. The second is if we colonise by way of terraforming, creating planetary ecosystems to support human and other forms of life. Even if the number of animals taken to begin such processes is small, creation of any flourishing ecosystem is going to very quickly lead to a large number of animals.

It is also possible that the future will not be dominated by either humans or non-human animals but digital beings—sentient AIs. In the end, there is a lot of uncertainty here and

[3] Current attempts to weight based on neuron count (e.g. MacAskill 2022) are unconvincing (see Shriver 2022).

unless we are quite sure of these alternative outcomes, we still have reason to believe that there will be very high numbers of animals in the future.

2.2 Suffering

As well as there being lots of animals (both now, and expected in the future), many of these animals will have bad lives. In the words of Beckstead there are: 'an astronomical number of expected future beings with lives that are suboptimal, and a future whose trajectory is potentially influenceable' (Beckstead 2019: 92). Though he was talking about pessimistic estimates of the lives of future humans, the same applies even more strongly for animals. There is thus a great amount of future suffering that we can potentially prevent.

From the numbers we presented above, we can see that almost 75% of land vertebrates live in agricultural systems. These systems are well known for the suffering caused to the animals (Harrison 1964; Singer 1975; Gruen 2011). Most broiler chickens spend their lives in windowless sheds with under one square foot per bird; their beaks are trimmed using hot blades to decrease the aggression brought on by the crowded conditions. They frequently suffer leg deformities and lameness from ongoing selective breeding for rapid growth. Sows used for breeding are often kept in tiny stalls in which they are unable even to turn around, with few cognitive or behavioural challenges/opportunities and no access to nesting materials to fulfil their strong drive for nest-building. For many, if not most, of these animals, there are almost certainly ongoing negative experiences and few opportunities for positive experiences such that their lives are highly likely to contain more suffering than pleasure. If current agricultural practices were to continue like this into the future, there would be ongoing suffering at a large scale. Again, one may counter that we should not expect high levels of future animal suffering simply based on current circumstances. If factory farming is going to end, or if conditions are going to vastly improve, then we will not have future suffering of food animals. As we will argue in what follows, even if this is true we may still see huge benefit in speeding up the trajectory.

Many wild animals also suffer. Many writers argue that, in fact, suffering dominates in nature (Ng 1995; Horta 2010; Tomasik 2015; Iglesias 2018). This is in part attributed to the general causes of suffering, such as injury, disease, starvation, and predation. However, it is also considered to be an effect of the life history of many wild animals—the 'r-selected' species that produce a large number of small or 'cheap' offspring, of which only a few live to maturity. The large numbers that instead perish are considered to have lives almost completely composed of suffering (from whatever processes kill them), with few if any opportunities for pleasure. Given the large numbers of such individuals, it is then taken to be the case that there is an overwhelming prevalence of suffering over pleasure. Though we think there are reasons to doubt that animal suffering in the wild outweighs positive experience (Browning and Veit 2023), it is obvious that it is still widespread. Overall, not only are there lots of animals, but they potentially have lives containing a lot of suffering, and that we can change for the better. Animal suffering is a major, if not *the* major, source of current disvalue, and plausibly so too in the long-term future. It should thus be accounted accordingly.

3 Potential interventions

We have argued here that it is important to include animals in calculations about which actions we should prioritise for the long term; which, due to the numbers and degree of suffering involved, is likely to lead to giving priority to animal-based actions in many cases. We do not rule out that in some (or even most) cases, the calculations will still favour human-centred interventions, for a range of reasons such as tractability or differential moral weight, but this should not be taken for granted without further investigation. Instead, relevant animal-based actions should also be assessed and compared. How, then, can this be done? In this section we will survey a number of different potential interventions that may improve the long-term future for animals. We don't make any strong claims about which would in fact be the best options to pursue, but discuss what we take to be some of the more promising avenues for further investigation.

The important categories for action in shaping the far future can be divided into 'proximate benefits', speeding up development, and trajectory changes (Beckstead 2019). Proximate benefits are the more predictable, short-term benefits of action. Speeding up development refers to pushing developments that could improve future quality of life to earlier in the timeline, such that their benefits will be felt for longer. Trajectory changes are arguably the most impactful, and involve shifting the direction of the world's development, such that we end up with a different kind of future than we otherwise would have had—an example of this being the abolition of slavery (MacAskill 2022). A current example relevant to animals may be the rise of global aquaculture—this is an industry that is still young, and the structures and regulations we set in place now may have long-reaching effects in terms of how the industry develops (Franks, Ewell, and Jacquet 2021).

One potential objection is that actions to improve human welfare may be the right priority right *now*, as ensuring the welfare of humans is also the best way to create a future in which animals are taken care of (Sebo 2022). Improving our social, economic, and political systems can help empower future generations and create space for developing a capacity for and desire to help animals. Not until our own needs are met can we perhaps then turn to assisting others. This also relates to a possibility that if most of the value in the far future will actually be realised by sentient AIs, created by humans, then we should be prioritising actions to protect humans and ensure the development and spread of such AIs. While we see value in this objection, it is not one that can just be asserted *a priori*. It may very well turn out to be the case as a result of our calculations, when we place both human and non-human animal wellbeing into the calculus. Importantly though, this decision needs to be made after making an assessment including all these factors, and with comparison to the set of possible alternative actions focusing more directly on animals. We are not claiming that these assessments would end up showing we should prioritise interventions to help animals, but merely that without including consideration of animal interests, we couldn't know for sure. We should also be wary of potential motivated reasoning toward conclusions that support our own self-interested preferences, and making sure to include reasoned considerations of non-human animal interests will help prevent this.

There are of course many different possible actions that could help improve the situation for animals in the long-term future, but here we will outline a few that are likely to be beneficial and are worthy of further investigation. They can be grouped into two categories—those

that change the number of future animals, and those that change the quality of life of future animals (changing the size of the future and changing its sign). That is, we should try to ensure there are lots of future animals if we predict their lives to be good, and few if we predict them to be bad. Additionally, we should work to try to improve expected quality of life such that all the animals who will exist will have lives of the highest positive welfare we can achieve. We take it that in response to observed suffering it is preferable to act to reduce the sources of suffering rather than the number of bearers of suffering (Višak 2017), wherever the former is possible. But which of these interventions different actors prefer will depend strongly on their ethical and axiological commitments.

3.1 Changing the number of animals

There are two ways in which we can beneficially change the size of the future regarding animals—one is in reducing the number of animals if we expect them to have bad lives, and the other is increasing the number of animals we expect to have good lives. This may be a complex question to answer in practice, as the differences in animal cognitive sophistication, lifestyles, and evolutionary history will influence their overall lifetime welfare balance—for example, prey animals may experience more fear from predator presence, while predators may be more stressed by the demands for successfully finding and hunting prey. However, in thinking about setting up a long-term future that contains few suffering animals and abundant happy animals it is important to think about which animals will have good or bad lives. This differs from changing the quality of animal lives from negative to positive as what we're considering is decisions about whether or not to bring animals into existence rather than how to make their lives better.

For the first—reducing the number of unhappy animals—one potentially important intervention is ending factory farming. Though the numbers involved are lower than for wild animals, the suffering is arguably higher—with most animals probably having strongly net-negative lives—and this is a more obviously tractable intervention than many of those discussed for wild animals. Thus, ceasing to bring animals into these situations would be a significant change to overall value, and a long-term future in which such practices no longer exist will be a far better one than if they do. For instance, widespread adoption of a vegan diet would lead to fewer animals used, and thus fewer numbers over time. Development of in vitro 'clean meat' products is one possible path to this end (Anomaly et al. 2024), as are general advocacy movements to increase veganism. In general, intensive farming is benefitted through direct subsidies and by externalising the costs of harms to health, environment, and animal welfare; simply altering these would make the industry far less economically viable (Sebo 2022). This is an example of where longtermist and short-termist goals align—reducing the number of suffering animals *now* and preventing far more being created in the future. Additionally, if, as John and Sebo (2020) argue, the existence of animal agriculture maintains human attitudes toward animals that hinder moral circle expansion, then elimination of this practice is crucial to ensuring the ongoing wellbeing of animals in the long-term future.

This could be considered a version of 'speeding up progress', if we think that animal agriculture will eventually die off, but that the sooner we reduce it, the more animals will be saved from coming into existence in a life of suffering. Given the large numbers and

suffering involved every year that intensive farming persists, any action we can take to bring this sooner will still represent a large gain. We can also here include possible additional benefits, such as reducing risks of future pandemics, and slowing down climate change. As will be discussed further on, it might also be considered as a movement toward a better attractor state. If we think that at some point the dietary preferences of humans will become fairly fixed in one state or the other, then pushing toward the higher-value state would ensure a better future. As we have already mentioned, we may resist this as a longtermist priority if we think that we are already on this path, such that factory farming is likely to end in the short- to medium-term future. Evidence for this could be seen in the increased adoption of a vegan diet (Russell 2023), and growing concern for the welfare of farmed animals; but on the other side we can see numbers of intensively farmed animals still continually increasing (Torrella 2021). While the population share of vegans increases, so too does the total number of humans who aren't. A lot here depends on where we see the current trend heading, and whether or not intervention now is needed to ensure this state in the long-term future. If we are at all uncertain about this trajectory, actions to ensure we bring about the more positive outcome would have a high expected value.

Another way in which we could reduce the number of suffering animals could be in reducing the number of wild animals, or at least those of the types we take to have lives predominantly composed of suffering. This seems to be the view taken by some writers (Tomasik 2017), who follow the 'logic of the logger' (John and Sebo 2020) in arguing that reduction of suffering entails habitat destruction, to decrease the number of animals. Ensuring a future with fewer or no suffering animals will increase its expected value. It is an open question as to how much of a current priority this should be, based on which specific actions now are likely to have uniquely strong effects on the numbers of wild animals in the far future. We take this to only be desirable if we are unable to instead intervene and improve the lives of these animals, an option we will discuss in the next section. While the former would reduce the amount of disvalue, the latter would also increase the amount of value, which will bring greater overall benefit.

The other way in which we can positively impact the size of the future is in ensuring there are large numbers of animals with positive welfare. The far future will have far greater total value with a large number of happy animals in existence than if suffering animals are simply absent. If it is the case that the existence of more net-positive lives is worthwhile, then we should be looking at ways to maximise the number of happy animals. One version of this would be mitigating extinction risk, at least for species with good lives. Like the mitigation of human extinction risk, this would allow for a future filled with much larger numbers of happy beings. Many of the efforts to mitigate extinction risk will align with those used for humans (e.g. addressing climate change, reducing the chance of meteor collision), but there will be some unique to animals. For many of these actions there will be complex trade-off calculations necessary, as resource distribution considerations require that increasing numbers of some species will place limits on others.

Prioritising the creation of animals who would have good lives would involve determining which animals may be capable of the most pleasure, and the conditions under which they should be kept to realise it, then investing resources in their creation and management. Understanding the relative sentience of different creatures, as well as allowing us to assess their relative suffering, will also give us guidance on what sorts of creatures we should be creating—which provide the most potential 'welfare per unit', so to speak.

If, for example, dogs are capable of as much pleasure as humans, but it is much simpler to provide them with what they need to achieve it, this gives us reason to promote the future numbers of dogs over those of humans. Depending on the empirical facts about relative wellbeing, it may even turn out that it is worth sacrificing potential numbers of humans in order to make this happen. This would also require understanding of the longer-term effects and side-effects—both positive and negative—of population expansions to judge the relative benefit.

3.2 Improving the lives of animals

As well as changing the size of the future we can also aim to change its sign—that is, to reduce suffering and increase pleasure for those animals who will exist. This could be done for agricultural animals and/or wild animals, or for the additional human-created animals described in the previous section. Reduction or elimination of animal agriculture is important where we think that animals in these conditions have net-negative lives, which is highly plausible for most modern practices. There is, however, also the possibility of changing farming practices such that animals have net-positive lives. In this case, the so-called 'logic of the larder' (John and Sebo 2020) would then advocate their creation and consumption, because the creation of positive lives is an overall good and should be encouraged. However, this is unlikely to have the expected benefits. If what we wanted to do was create the greatest number of happy animal lives, it is doubtful that agriculture would be the most cost-effective way to do so. We could, for example, raise large colonies of happy mice for far less money than the agricultural industry takes to sustain, as well as freeing up cropland currently used to feed agricultural animals, providing habitat for more wild animals (Matheny and Chan 2005). There are additionally the potential negative societal effects of animal consumption, particularly in terms of poor human attitudes toward animals leading to poor welfare outcomes overall (John and Sebo 2020).

With wild animals being probably the second highest source of animal suffering—higher in numbers but with more opportunities for positive welfare experiences to offset their suffering—investigating ways to manage wild animals to remove many of the negative experiences is another research priority. As mentioned, some advocate for the reduction in wild animal numbers as the best way to reduce suffering (Tomasik 2017), however, if it is possible instead to switch net-negative to net-positive lives, this will be a superior intervention than simply removing such lives. Currently, discussions of intervening on wild animal welfare are hampered by the sheer complexity of the task—we are famously terrible at making ecosystem changes without hosts of downstream negative effects. However, the more we know, the more possible it will be to do so, and perhaps aiming for a future in which we have the knowledge and ability to manage all wild animal populations for their maximal welfare would be ideal.

Another method would be to ensure that all the animals who do exist—captive or wild—are capable of increased wellbeing, not just through better life conditions but through use of technologies that make these animals capable of experiencing more pleasure (and/or less suffering). This could include selectively breeding or genetically engineering animals to

have a greater capacity for total pleasure, and/or an ability to take more pleasure in the conditions under which they usually find themselves. Some possible methods to achieve this include carefully managed gene-drives to introduce and spread genes that enhance welfare (Liedholm 2019), use of enhancement drugs that increase pleasure or take away suffering—discussed in the human case (Veit 2018; Veit et. al 2020) but so far given little attention for animals—or even development of technology for applying stimulation to the pleasure centres in the brain, which does not appear to be subject to diminishing marginal utility as other types of pleasures do (Ng 1997).

Related to this could also be engineering animals for reduced suffering. There is a small but growing literature regarding use of genetic engineering to create so-called 'diminished' animals who lack some of the species-typical capacities that currently create frustration and suffering, including the most extreme case 'animal microencephalic lumps' that completely lack sentience (Schultz-Bergin 2017). However, as this method removes the possibility of good lives and positive values, it will not end up creating the highest expected utility except in cases where the suffering would otherwise be inevitable. Where there are opportunity costs of directing resources away from creating or supporting otherwise happy animals, this would reduce total value.

These types of intervention may also interact with the changing animal numbers. Say, in the future, we are capable of creating very happy animals through use of chemical intervention or genetic engineering—this would then give us a reason to try to create and maintain as many animals as possible that are capable of experiencing this. One suggestion resulting from this is that this then might give us reason to try to maintain factory farms, as these are capable of holding the highest densities of animals, if the suffering currently experienced in such setups would be replaced with the types of pleasure described. However, this would only be true if we took factory farms as the best way of housing and keeping happy animals. This may be the case if we think that the economic incentives of using animal products would offset the costs of animal maintenance, but as we have discussed, it is also likely that there are many other setups and housing types that would, in actuality, be better for keeping large numbers of happy animals.

Some of these interventions may not be appealing to those with a less utilitarian approach to animal ethics, who have more of a concern for other values in animal lives, such as authenticity, or naturalness. We think there are reasons to question the role of such values: for instance, there is no obvious link between naturalness and moral value (Browning 2020)—pain, suffering, and extinction are, after all, perfectly natural phenomena and yet it is precisely these that we would most wish to avoid. It is important in any view of animal ethics to avoid anthropocentrism by considering what the animals themselves care about rather than imposing a human-centred view of what is valuable. Whereas humans might object to the prospects of being prescribed mood-enhancing drugs on grounds of authenticity, autonomy, or consent (see Veit 2018), there is less reason to think that such reasons would apply to animals. For example, as we have argued elsewhere, there is little reason to think that freedom must matter intrinsically to animal welfare (Browning and Veit 2020; 2021). Regardless, it is not our intention here to take a strong stance on views in animal ethics, and we take it to be the case that even where one is not in favour of some of these specific utilitarian actions, this still leaves a range of plausible interventions that undeniably improve animal lives on any account.

3.3 Value change

We have described a number of possible actions for improving the long-term future for animals, in terms of changing both the size and the sign of this future. However, more consideration is needed regarding which are likely to be the most effective actions for future benefit. When thinking about the long-term future, effective interventions will be those that persist for a long time, and are robust in the face of potential changes in future conditions. This can be framed in terms of attractor states—those states of the world that, once entered, are likely to continue for a long period (Greaves and MacAskill 2019). There are many potential attractor states, and some will be of more value than others. If our actions now can affect the probability that we enter a better rather than a worse attractor state, this will have ongoing effects. If we want to take the best possible actions for long-term future value, then focusing on the best attractor states is most likely to have the highest ongoing value. One potential action that we think may serve as an attractor state for the future of animals is change in human values and attitudes towards animals and their treatment, in terms of both individual and larger-scale institutional values.

Interventions to alter human attitudes and values could form a type of trajectory change from one attractor state to another, such that future policy and behaviour will be different. We can plausibly influence the direction of individual and institutional values toward those most likely to have positive future impacts for animals. This would include any action to ensure that future humans, particularly those with political power, hold attitudes that promote positive treatment of animals. Such changes will potentially have wide-ranging effects across all the domains we have described, ensuring their implementation and maintenance.

If our current stage in time is a particularly notable time in which we are about to see some form of 'value lock-in' (MacAskill 2020; 2022), where previously flexible or pluralist values give way to a single or rigid set of values that persist over a long timescale, then it is worth channelling resources now to ensure that these values are those that ensure good lives for future animals (and humans). This may be particularly likely if we think we are on the cusp of programming future superintelligent or autonomous AIs that will have a large influence on politics and society (Bostrom 2014). The values with which we program such AI systems can affect animals as much as they do humans, and the values they receive now may have a strong future influence on the conditions of animal lives. Though some scepticism has been expressed regarding how easily we may accurately represent animal interests (Ziesche 2021), even just ensuring that non-speciesist and 'animal friendly' attitudes are included should help ensure animals are given appropriate consideration.

One form such an attitude change could take would be a moral circle expansion—i.e. widening the circle of beings recognised as subjects of moral concern. This could include sentient animals and, potentially, other beings such as sentient AIs, which have recently received a surge of attention in the longtermist literature (John and Sebo 2020). Previous moral circle expansions based on shared humanity have provided increased protections for marginalised groups, and further expansion based on sentience would provide protection for animals, including farmed and wild animals (Anthis and Paez 2021). Moral circle expansion is compatible with a range of ethical frameworks; all it requires is that the type

of consideration that is currently offered only to humans would also be offered to other sentient animals. In practice this should mean that animal suffering counts for much more than it currently does; and many harmful practices could not continue. Depending on whether or not our current trajectory is already leading us to such an expansion, we could aim either to direct or to speed up expansion, both of which can have long-term benefits (Anthis and Paez 2021).

Even where one takes most of the future value to instead be realised by digital minds, it could be argued that this provides a reason to take animals seriously now, as neglecting animal interests may work to lock in anthropocentric values that would make it impossible in the future to ensure the consideration of the interests of other types of non-humans. It is possible that extending our moral circle to include other sentient animals may be a prerequisite for further extension towards sentient AI, the moral value of which is still not widely considered by either philosophers or the public. Such moral circle expansion, to include all sentient beings, would also require increased research into sentience and its neurological, cognitive, and functionalist basis to settle questions about which animals and AIs are sentient, and what their experiences are like. In light of the potential that our present is a particularly influential time where there is a higher risk of locking in harmful values into our institutions and legal frameworks, particularly when considering current work with AI, there may be a matter of urgency for pushing moral circle expansion right now, to ensure that it happens at all. Making sure that animal interests are included will help guard against cementing existing anthropocentric biases.

A concern about interventions based in human attitude change is whether or not they would actually work to improve the situation for animals. For instance, some take the historical evidence to speak *against* a correlation between value change and welfare—while we are arguably living in a time in which we hold the most animal-positive views, worldwide animal welfare is at its worst. However, this take misrepresents the current situation, and the relevant comparison class. The worsening state of animal welfare is largely due to increasing numbers being held in factory farms, which is a function of increasing population size and the spread of intensive farming techniques into new populations (Torrella 2021). If we hold fixed this increase in population size and look at the counterfactual situation regarding values—one in which societal attitudes towards animals had remained largely fixed, or gotten worse—then it is highly likely that the current situation would be even worse than it is now. It seems like we are closer than ever before to ending the practice of intensive farming, and human value change has been leading changes in practices, such as increasing adoption of bans on veal crates, sow stalls, and battery cages for chickens, that are an overall net benefit for welfare.

What is important is that we examine the causal links between societal values and the conditions of animal housing and husbandry. This will allow us to determine where best to target our interventions to create lasting and relevant value change. Whether individual changes of attitudes or wider structural change and improved institutional decision-making is more important may depend on what are the dominant mechanisms for value change, which is a matter for further research (see e.g. Sebo 2022). These should also not be taken as exclusive options, and indeed will often be complementary in that critical mass of individual lobbying or purchasing decisions, as well as research and policy advice, will influence institutional change.

4 Conclusion

The longtermist paradigm holds that the actions expected to produce the most good are those that have their effects in the long-term future. Here, we have argued that for the same reasons this argument is applied to considerations regarding humans, the wellbeing of future animals should also be given serious consideration when thinking about the long-term future. In fact, there is an interesting parallel between general longtermist thinking and an emphasis on the importance of animal welfare. Both are situations in which the group concerned (far-future populations or animal populations) compose a vast majority yet their interests are subsumed to the interests of a small majority (i.e. humans, or members of current and nearby future generations) and where the individuals concerned lack any political representation for their interests. There is thus a natural alignment between longtermism and more traditional animal advocacy, with promise for further collaboration. Studying the methods for creating successful change for consideration of the interests of animals who cannot make their voices heard may help us to likewise influence political institutions to take future humans, animals, and sentient AIs into account.

As well as including animals within longtermist thinking, we should additionally recognise the possibility that in some cases their aggregate interests may even dominate, due to their greater numbers and greater possible suffering. Even if one wants to resist this and maintain an anthropocentric priority, it is clear we should be giving animals much more consideration than is currently the case. We have not attempted to quantify the size of these effects or look at the relative calculations of the expected value contained in future animal versus human lives, but we suspect this could support an even stronger conclusion. That is, given the reasons we have presented, we actually have greater reasons to consider animals than humans: that our *best* actions in future will be those which benefit animals. Given the sheer numbers and level of suffering we see right now, it could even turn out to be the case that many short-term interventions to benefit animals could be more valuable than long-term interventions for human societies: for instance the lower bound estimate of a total of 10^{15} future people (Greaves and MacAskill 2019) is equal to the number of aquatic vertebrates existing *right now*. However, given the uncertainty surrounding the actual future numbers and level of suffering, as well as the comparative moral weight to humans, here we will content ourselves with the weaker claim that animals should at least be brought into deliberations regarding our best actions for the long-term future with a much greater weight than they are currently accorded. This is likely to significantly change the landscape of action prioritisation for the long-term future. We have described a range of potential interventions that change both the expected size and sign of the future, highlighting that actions targeted at changing human and societal attitudes are most likely to have a strong effect.

In some cases, the actions we have described (such as ending factory farming) will align with short-term priorities, but most often they are likely to focus on different initiatives—ending the most animal suffering *now* is not necessarily related to ending it in the long term, as is also true for human cases. Here, we sometimes have to push against our intuitions that we should be doing something now, if we accept the motivations for a longtermist worldview. Indeed, if we do not think we are living in a particularly influential period in time (i.e. one in which our interventions are likely to have unusually strong ongoing effects), then it may in fact be better to invest our resources such that they can be used for future interventions, rather than to take any direct action now (MacAskill 2020). Particularly where we are

currently uncertain about the specific changes that may end up being best for animals in the long run, empowering future generations to act on their behalf through shifting values and building up knowledge will be our current best action. However, some of the most important interventions discussed—such as institutional change and moral circle expansion, will have immediate as well as long-term effects.

The upshot of this chapter is not to advocate some specific action/s, but to call for the inclusion of animals in deliberations about the long-term future and which actions we should be prioritising for greatest gain. Importantly, it is a call for further research. While these questions are still uncertain, we should be gathering information such as the impact of attitude change on future behaviour, the net balance of pleasure/suffering in wild animals, and the likely future numbers of animals and humans. The mere assertion that we have little knowledge about how to improve the lives of animals is not enough to exclude them from a longtermist view, since it is precisely here that all future knowledge about animal welfare should be included. By bringing animals into our considerations, we can be surer that we will be making decisions that will have the best actual long-term impact, and it is our hope here that we have shown that the interests of animals should play a much larger role in longtermist thought and writing.

References

Anomaly, J., Browning, H., Fleischman, D., and Veit, W. (2024), 'Flesh without blood: The public health benefits of lab-grown meat', *Journal of Bioethical Inquiry*, 21(1), 167–175.

Anthis, J. R. and Paez, E. (2021), 'Moral Circle Expansion: A Promising Strategy to Impact the Far Future', in *Futures* 130: 102756.

Bar-On, Y. M., Phillips, R., and Milo, R. (2018), 'The Biomass Distribution on Earth', in *Proceedings of the National Academy of Sciences* 115/25: 6506–6511.

Beckstead, N. (2019), 'A Brief Argument for the Overwhelming Importance of Shaping the Far Future', in H. Greaves and T. Pummer (eds.), *Effective Altruism: Philosophical Issues* (Oxford University Press), 80–98.

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', in *Utilitas* 15/3: 308–314.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4/1: 15–31.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

Browning, H. (2020), 'The Natural Behavior Debate: Two Conceptions of Animal Welfare', in *Journal of Applied Animal Welfare Science* 23/3: 325–337.

Browning, H. (2023), 'Welfare Comparisons Within and Across Species', in *Philosophical Studies* 180: 529–551.

Browning, H. and Veit, W. (2020), 'Confined Freedom and Free Confinement: The Ethics of Captivity in *Life of Pi*', in Á. T. Bogár and R. S. Szigethy (eds.), *Critical Insights: Life of Pi* (Salem Press), 119–134.

Browning, H. and Veit, W. (2021), 'Freedom and Animal Welfare', in *Animals* 11/4: 1148.

Browning, H. and Veit, W. (2023), 'Positive Wild Animal Welfare', in *Biology and Philosophy* 38/14: 1–19.

fishcount.org.uk. (2019), 'Fish Count Estimates', *fishcount.org.uk*, http://fishcount.org.uk/fish-count-estimates-2(accessed 25 February 2021).

Franks, B., Ewell, C., and Jacquet, J. (2021), 'Animal Welfare Risks of Global Aquaculture', in *Science Advances* 7/14: eabg0677.

Greaves, H. and MacAskill, W. (2019), 'The Case for Strong Longtermism', GPI Working Paper No. 7-2019 (Global Priorities Institute, Oxford University), https://static1.squarespace.com/static/5506078de4b02d88372eee4e/t/5f1704905c33720e61cd3214/1595344019788/The_Case_for_Strong_Longtermism.pdf (accessed 25 February 2021).

Gruen, L. (2011), *Ethics and Animals: An Introduction* (Cambridge University Press).

Harrison, R. (1964), *Animal Machines: The New Factory Farming Industry* (Vincent Stuart).

Horta, O. (2010), 'Debunking the Idyllic View of Natural Processes: Population Dynamics and Suffering in the Wild', in *Télos* 17/1: 73–88.

Iglesias, A. V. (2018), 'The Overwhelming Prevalence of Suffering in Nature', in *Revista de Bioética y Derecho* 42: 181–195.

John, T. and MacAskill, W. (2021), 'Longtermist Institutional Reform', in N. Cargill and T. John (eds.), *The Long View* (FIRST), 44–60.

John, T., and Sebo, J. (2020), 'Consequentialism and Nonhuman Animals', in D. W. Portmore (ed.), *Oxford Handbook of Consequentialism* (Oxford University Press), 564–591.

Liedholm, S. E. (2019), *Persistence and Reversibility* (Wild Animal Initiative), https://static1.squarespace.com/static/5f04bd57a1c21d767782adb8/t/5f160c91bc0bff4abe964d5a/1595280529848/WAI_PersistenceAndReversibility_Dec2019.pdf (accessed 25 May 2021).

MacAskill, W. (2020), 'Are We Living at the Hinge of History?', GPI Working Paper No. 12-2020 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/wp-content/uploads/William-MacAskill_Are-we-living-at-the-hinge-of-history.pdf (accessed 15 March 2021).

MacAskill, W. (2022), *What We Owe the Future* (Hachette).

Matheny, G. and Chan, K. M. A. (2005), 'Human Diets and Animal Welfare: The Illogic of the Larder', in *Journal of Agricultural and Environmental Ethics* 18/6: 579–594.

Mogensen, A. (n.d), 'Staking Our Future: Deontic Longtermism and the Non-identity Problem', GPI Working Paper Series (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/wp-content/uploads/2019/Mogensen_Staking_Our_Future.pdf (accessed 15 March 2021).

Ng, Y. K. (1995), 'Towards Welfare Biology: Evolutionary Economics of Animal Consciousness and Suffering', in *Biology and Philosophy* 10/3: 255–285.

Ng, Y. K. (1997), 'A Case for Happiness, Cardinalism, and Interpersonal Comparability', in *The Economic Journal* 107/445: 1848–1858.

Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).

Russell, A. (2023), 'How Many Vegans Are There in the World?', *WTVox*, https://wtvox.com/lifestyle/2019-the-world-of-vegan-but-how-many-vegans-are-in-the-world/ (accessed 30 November 2022).

Schultz-Bergin, M. (2017), 'The Dignity of Diminished Animals: Species Norms and Engineering to Improve Welfare', in *Ethical Theory and Moral Practice* 20/4: 843–856.

Sebo, J. (2022), *Saving Animals, Saving Ourselves* (Oxford University Press).

Shriver, A. (2022), 'Why Neuron Counts Shouldn't Be Used as Proxies for Moral Weight', *EA Forum*, https://rethinkpriorities.org/publications/why-neuron-counts-shouldnt-be-used-as-proxies-for-moral-weight(accessed 30 November 2022).

Šimčikas, S. (2020), 'Estimates of Global Captive Vertebrate Numbers', *Rethink Priorities*, https://rethinkpriorities.org/research-area/estimates-of-global-captive-vertebrate-numbers/ (accessed 30 November 2022).

Singer, P. (1975), *Animal Liberation* (Harper Collins).

Tarsney, C. J. (2020), 'The Epistemic Challenge to Longtermism', working paper (PhilPapers), https://philpapers.org/archive/TARTEC-2.pdf (accessed 15 March 2021).

Thorstad, D. and Mogensen, A. (2020), 'Heuristics for Clueless Agents: How to Get Away with Ignoring What Matters Most in Ordinary Decision-Making', GPI Working Paper No. 2-2020 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/wp-content/uploads/David-Thorstad-Andreas-Mogensen-Heuristics-for-clueless-agents.pdf (accessed 15 March 2021).

Tomasik, B. (2015), 'The Importance of Wild-Animal Suffering', in *Relations: Beyond Anthropocentrism* 3/2: 133.

Tomasik, B. (2017), 'Habitat Loss, Not Preservation, Generally Reduces Wild-Animal Suffering', *Essays on Reducing Suffering*, http://reducing-suffering.org/habitat-loss-not-preservation-generally-reduces-wild-animal-suffering/ (accessed 26 March 2021).

Torrella, K. (2021), 'The Biggest Animal Welfare Success of the Past 6 Years, in One Chart', *Vox*, https://www.vox.com/future-perfect/22331708/eggs-cages-chickens-hens-meat-poultry (accessed 30 November 2022).

Veit, W. (2018), 'Cognitive Enhancement and the Threat of Inequality', in *Journal of Cognitive Enhancement* 2: 404–410.

Veit, W., Earp, B. D., Faber, N., Bostrom, N., Caouette, J., Mannino, A., Caviola, L., Sandberg, A., and Savulescu, J. (2020), 'Recognizing the Diversity of Cognitive Enhancements', in *AJOB Neuroscience* 11/4: 250–253.

Višak, T. (2017), 'Preventing the Suffering of Free-Living Animals: Should Animal Advocates Begin the Killing?', in *Journal of Animal Ethics* 7/1: 78–95.

Ziesche, S. (2021), 'AI Ethics and Value Alignment for Nonhuman Animals', in *Philosophies* 6/2: 31.

PART 4
INSTITUTIONS AND SOCIETY

26

Longtermist Political Philosophy

An Agenda for Future Research

Andreas T. Schmidt and Jacob Barrett

1 Introduction

What we do now can affect future generations. And most people agree that, morally, we ought to consider their interests. But how far into the future do our obligations reach? Political debates typically focus on the near- to medium-term. For example, even discussions of climate change often focus only on the next century. But while most CO_2 emitted today will leave our atmosphere within 200 years, the remaining 20–35% could remain for thousands of years longer (Emanuel 2016: 15). Shouldn't we care about such longer-term effects, too?

Longtermists answer that we should. They argue that we should give significant moral weight to positively affecting the distant future, and that this has important and fascinating implications for what we should do now.[1] The standard case for longtermism goes roughly as follows.

First, longtermists argue that distant future people matter morally. Economists and policymakers commonly discount future income or consumption. However, most philosophers argue against discounting non-instrumentally valuable things, such as well-being, simply because they lie in the future.[2] Future well-being should not be subject to a positive rate of pure time preference, or the rate should at least be small.[3] This still allows that we might discount future people's well-being for other reasons.[4] For example, Tyler Cowen suggests that our annual risk of extinction is high enough that we should mainly consider people in the next few thousand years, and heavily discount the well-being of those living in the more distant future since they are unlikely to exist.[5] Call a view focused on such time scales 'medium longtermism'. Other longtermists have longer time horizons and believe we should focus on people millions or billions of years in the future. Such 'full longtermism' may follow from more optimistic estimates of our chance of surviving that long, or from the

[1] See, e.g., Cowen (2018), Ord (2020), MacAskill (2022), Greaves and MacAskill (this volume).

[2] See, e.g., Cowen and Parfit (1992), Broome (1994), and the citations at fn. 21 in Greaves and MacAskill (this volume). See also Heath (2017) and Russell (2022) for opposing arguments.

[3] Arguments for adopting a small discount rate (which is too low to challenge longtermism) appeal to factors such as agent-relative reasons (Mogensen 2022) and uncertainty about the right discount rate (Weitzman 1998).

[4] As we will see, some also downgrade the moral significance of distant future people's well-being by drawing a sharp moral distinction between existing and future people generally, for example, by adopting 'person-affecting' views in population ethics.

[5] Cowen (2023) lays out this view during a talk on effective altruism, but it is also common to discount future people's interests for this reason, for example, in climate economic models (Stern 2006).

Andreas T. Schmidt and Jacob Barrett, *Longtermist Political Philosophy* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Andreas T. Schmidt and Jacob Barrett 2025.
DOI: 10.1093/9780191979972.003.0026

idea that even a small probability of this occurring is enough for very distant future people's well-being to weigh heavily *in aggregate*, given the sheer number who might exist.

This brings us to a second major longtermist claim: there might be *vast* numbers of far-future people. For example, Greaves and MacAskill argue that a plausible estimate of the expected number of future people, conditional on humanity remaining Earthbound, is 10^{14}—that is 10,000 persons for every person alive today. They see this as a lower bound estimate, noting that if we assign even small probabilities to humanity settling the stars, or to future sentient life being digital, the number of expected future people magnifies greatly: here their estimates range from 10^{18} to 10^{45} future people (Greaves and MacAskill this volume).[6] Exact numbers are neither possible nor necessary, and we should treat such estimates with a heavy grain of salt. But in any case, the expected number of future people is likely orders of magnitude greater than people alive today, such that the value at stake in the long-term future may be massive.[7]

Third, longtermists argue that we can sometimes predictably affect the value of the long-term future. Most obviously, we can try to mitigate extinction risks that threaten to massively curtail far-future value by causing our extinction. These might be, for example, risks due to nuclear war, pandemics, climate change, or artificial intelligence (Ord 2020). Alternatively, we might also be able to affect the value of the long-term future, conditional on humanity's survival, by launching ourselves on higher-value trajectories (MacAskill 2022): for example, those less ravished by climate change, repressive governments, or catastrophic events so terrible that humanity can't bounce back.[8] Note also that *if* such ways of improving the distant future exist, it would be surprising if we had already identified and were already pursuing all of them. After all, longtermist thinking is not mainstream and, unlike most other groups, future people cannot advocate for their own interests, for example, through participating in politics or markets.

These three ideas suggest that our actions can, in expectation, sometimes bring about a huge amount of value in the long run. If we also believe that we morally *ought* to promote this value—something we discuss in Section 2—then we arrive at longtermism. Informally stated for now, this is the thesis that positively affecting the value of the long-term future is *among* the key moral priorities (call this 'weak longtermism') or perhaps *the* key moral priority (call this 'strong longtermism') of our time (MacAskill 2022: 4).

Many discussions of longtermism focus on individuals—for example, on where philanthropists should donate. However, institutions also affect the long-term future. Longtermism may thus be an important topic for *political philosophy*: positively affecting the value of the long-term future may be a (or the) key *institutional* priority of our time.

[6] A common misconception is that longtermists endorse the *goal* of bringing sentient digital beings into existence. More standardly, longtermists appeal to sentient digital beings to emphasize the stakes of making the future go well because of how large the future population of sentient beings might be.

[7] Later, we flag views that drive a wedge between huge numbers of future people and huge moral stakes. Some might also deny the relevance of the *expected* future population size, given its dependency on tiny probabilities of massive populations, and focus only on more likely scenarios.

[8] Some longtermists use the broader category of an 'existential risk' to refer to any 'risk that threatens the destruction of humanity's long-term potential' (Ord 2020: 37). This is meant to include risks not only of extinction but also of other similarly significant and irreversible events, such as an irrecoverable civilization collapse or an advanced artificial intelligence (AI) permanently disempowering humanity. However, as the distinction between (non-extinction) existential risks and trajectory changes is fuzzy and often causes confusion, we avoid talk of 'existential risks' here.

Now, political philosophers have produced valuable work on future generations, involving climate change, environmental ethics, and intergenerational justice.[9] But these discussions often focus on the near term. Moreover, while longtermists often invoke consequentialist or beneficence-based reasoning, political philosophers focus more on non-consequentialist considerations and on political values like justice and legitimacy. So, to date, there is almost no political philosophy explicitly engaging with longtermist thought. This is a striking lacuna, as longtermism might radically challenge mainstream political philosophy, and, conversely, political philosophy might offer challenges or insights for longtermists.

In this chapter, we outline some central questions in *longtermist political philosophy*. Our aim is to set out longtermist political philosophy as a research field and to motivate readers to pursue questions and arguments we here only broach. We believe more work in this area is both theoretically and practically important. Longtermism has spawned a thriving research field but also a flurry of philanthropic activity moving millions of dollars each year. It is important to understand what institutional implications longtermism might have, as well as how political values might constrain, challenge, or support longtermism (or particular longtermist proposals).

In section 2, we outline 'institutional longtermism', suggesting that the standard case for longtermism may be more robust for institutions than for individuals. In section 3, we turn to points of tension and convergence between longtermism and some values that loom large in mainstream political philosophy. Section 4 provides a grab bag of other questions we lack space to explore.

2 The case for institutional longtermism

We begin by distinguishing individual from institutional longtermism:

> *Individual longtermism*: when evaluating the moral choiceworthiness of individual actions or projects, we should often give significant weight to their expected long-term effects.

> *Institutional longtermism*: when evaluating the moral choiceworthiness of institutions, we should often give significant weight to their expected long-term effects.

The above definitions invoke 'significant weight' and thus remain neutral between the long-term future being *the* most important priority or it being *one* important priority. Implicitly, we are mostly concerned with the latter thesis of weak longtermism, though we assume all versions of longtermism are revisionary in recommending that we give much *more* weight to the long-term future than existing individuals and institutions tend to. The term 'often' is also intentionally vague and signals that longtermism may have revisionary implications in many, but not all, contexts. Finally, by 'institutions', we mean social practices with rule-governed behavioural interdependencies. These include both formal institutions, such as laws and public policies, and informal institutions, such as norms and conventions.

[9] See, e.g., Rawls (1971: sec. 44), English (1977), Barry (1997), Gosseries and Meyer (2009), Gardiner et al. (2010), Moellendorf (2015), González-Ricoy and Gosseries (2017).

Functioning formal and informal institutions often depend on and overlap with one another; for example, laws may command compliance only in the presence of supportive social norms (Barrett and Gaus 2020). However, for reasons of space, we primarily focus on formal institutions here.[10]

The standard argument for longtermism is beneficence-based. The basic idea is that our expected effects on far-future value are sometimes massive and sometimes much larger than our short-term effects, given that—as discussed above—our actions might affect a vast number of future people. To reach the deontic conclusion that we morally *ought* to promote long-run value, we could appeal to an all-things-considered *consequentialist* duty to do the most good. However, the standard argument requires only that we have a weighty *pro tanto* duty of beneficence to promote the good. This beneficence-based case can be made for both individual and institutional longtermism. For the latter, we can endorse *social beneficence* (or 'instrumentalism'): how much good institutions do is one central consideration when evaluating their choiceworthiness (Barrett 2022; Schmidt 2023a).

We discuss duties and values other than beneficence later. But, first, we consider four empirical reasons, and then four more 'philosophical' ones, why the beneficence-based case for longtermism might be more robust for institutional than for individual longtermism.[11] As this chapter outlines a research agenda, our discussion should not be read as providing fully worked-out arguments so much as ideas meriting further consideration.

2.1 Institutions may affect the long run

We first consider reasons why institutions can be impactful in ways that significantly, and in expectation, affect the value of the long-term future. These reasons suggest that, if you are already sympathetic to longtermism, you should care about institutions.

The first reason is the sheer *scale* of the power and influence of institutions. Consider states. The 2022 United States federal budget had roughly US$6 trillion in outlays, with yearly defence spending alone of $754 billion (Office of Management and Budget 2021). The Chinese Communist Party is arguably the most powerful organization in the world, ruling over roughly 1.4 billion people and influencing many more. Furthermore, government budgets underestimate the power of states, which also exert influence through legislation. And the *de jure* monopoly on violence (and its *de facto* implications) makes the government an actor like no other.

Second, institutions' impacts are *broad*. Good institutions are arguably *all-purpose goods* that are valuable when dealing with most large-scale societal challenges—such as pandemics, international crises, or climate change. What makes institutions 'good' depends on various factors, including collective epistemic qualities and decision-making procedures,

[10] Note two other features of our definitions. First, they are concerned with moral choiceworthiness, or what it would be morally better or worse to choose, factoring in all relevant moral considerations (both axiological and deontological). Second, they are concerned with *expected* rather than *actual* long-term effects. However, the term 'expected' is employed loosely, and is meant to allow, besides expected value theory, other approaches to decision-making that are risk- or ambiguity-averse, that discount tiny probabilities, and so on.

[11] To clarify, the arguments below are not meant to suggest that individual longtermists should focus their efforts on institutional change rather than on some other priority. We are concerned with whether institutional longtermism is more robust than individual longtermism, not with whether the implications of individual longtermism are distinctively institutional.

suitable conditions for collective action and public good provision, low corruption, and so on. Conversely, bad institutions can be 'all-purpose bads' (or at least can be 'all-purpose worse' than good institutions). For example, a kleptocratic system that extracts resources to benefit a small corrupt elite may not only be bad for its contemporaries, but may, in expectation, deal less well with major risks that come its way.

Third, institutions can sometimes have *long-term* impacts that, in expectation, affect the value of the long-term future. Some institutions are deliberately designed to have long-lasting effects (think of those targeting climate emissions) or to themselves be long-lasting (think of constitutions). But more generally, functional institutions tend to be sticky—or to 'reproduce themselves'—simply because they are functional. Social scientists talk of *path dependence*: past events, decisions, technologies, and institutions constrain later ones (Mahoney 2000; Pierson 2000). This can take drastic and far-reaching forms. The US Constitution still greatly influences and constrains decisions in the US today. Many former colonized countries still struggle to shake off the dysfunctional institutions colonizers installed. At the far end of the spectrum, work in 'persistence studies' uses explanatory variables hundreds or thousands of years old to explain social and economic outcomes today.[12]

Path-dependence means that institutions affect not only later outcomes and decisions, but also what institutions we come to have. And while we cannot precisely predict this, social science and history might uncover helpful heuristics. For example, Daron Acemoglu and James Robinson argue that countries sometimes display virtuous and vicious cycles in their long-term development (Acemoglu and Robinson 2012). Countries with inclusive economic systems often generate more inclusive political systems which in turn beget more inclusive economic systems (a virtuous cycle). Countries with 'extractive' economic systems might see organized extractive interests capture political influence, which in turn begets more extractive economic institutions (a vicious cycle).

A fourth consideration strengthens the others: under some scenarios, the scale, breadth, and stickiness of institutions may *intensify* in the future, as new technical capabilities and organizational forms increase the extent and persistence of institutional impact. Longtermists sometimes worry about *totalitarian risk*: a totalitarian government locking in its power over the long term. Bryan Caplan, for example, argues that future technological developments (such as AI-driven surveillance and life extension) might allow dictators to cement their power for a very long time (Caplan 2011). Alternatively, an artificial general intelligence (AGI) might become all (or very) powerful, such that the goals or values it receives at inception might get locked in long-term (MacAskill 2022: ch. 4). If that happens, it is crucial we have good institutions and values when this lock-in occurs.

Various empirical considerations thus seem to support institutional longtermism: institutions have large, broad, long-term, and perhaps intensifying impacts. But how long-term are we talking here? Does institutional longtermist lend itself more to medium or to full longtermism? In general, the most plausible way to have a *very* long-term impact is through effects on the near term that, with sufficient probability, persist for a very long time. This is why longtermists focus so much on extinction: if humanity goes extinct, this will almost certainly last forever. To the extent that institutions can reduce extinction risks, this may support full institutional longtermism. Furthermore, it might suggest an 'urgent' approach

[12] See, e.g., Giuliano and Nunn (2020), Abad and Maurer (2021), and especially Sevilla (2022) for critical reviews of such work.

to institutional longtermism focused on establishing institutions, as soon as possible, that can combat pressing risks arising from engineered pandemics, climate change, nuclear war, AI, and the like.

But how robust is this case for full longtermism? Consider two worries. First, the idea that preventing extinction has great long-term value goes through most easily on the assumption that it is better for the future to contain more (happy) lives. But some views in population ethics deny this, holding that the effects of human extinction, beyond its effects on those living today, would be morally neutral or even positive (Benatar 2008: ch. 6; see Thomas 2023 for discussion). Second, if the annual extinction risk is high, fending off such risks now might have little impact over the very long term: over the next few millennia, extinction might be close to guaranteed even if we prevent it the next century. Indeed, to have a seriously long-term impact through this route, full longtermists arguably must endorse the controversial *time of perils hypothesis*: if we make it through the next century or so—the 'time of perils'—the annual extinction risk will then reduce to a very low rate (Friederich and Aebischer 2021; Thorstad 2023).

Arguably, the case for medium longtermism requires fewer contested normative and empirical assumptions. For, even beyond their effects on extinction risk, some institutions have a reasonable prospect of improving the well-being of future people, conditional on our survival, over medium timelines. Think, for example, of institutions explicitly designed to represent future generations (González-Ricoy and Gosseries 2017; John and MacAskill 2021). Within such medium timelines, we might also promote value by engaging in 'patient' approaches to longtermist institutional reform, focused on trying to achieve good institutions *in* the long run, even if it takes a long time to get there. This doesn't necessarily mean that we should come up with a blueprint for institutions that we wish to achieve in the distant future, which we should aim at by trying to take steps from here to there—as 'ideal theorists' sometimes suggest (Rawls 2001; Simmons 2010). Instead, given our limited predictive capacities, we might better promote long-run institutional reform by trying to make institutions more *progressive*—that is, better at getting better—for example, by facilitating institutional experimentation and learning (Barrett 2020c; MacAskill 2022: 99–102).

So, the claim that institutions can significantly and positively affect the long-term future over medium longtermist time horizons seems plausible, but extending this to full longtermism requires a defence of more contested normative and empirical assumptions. However, since the same assumptions are needed to get full *individual* longtermism, institutional longtermism seems, so far, on at least as strong footing as individual longtermism.

2.2 Institutional longtermism may be morally robust

We have discussed how institutions might promote long-run value; we now consider how this impacts their choiceworthiness. As we have noted, the case for longtermism typically assumes a *pro tanto* duty of (social) beneficence. While many worries have been raised about this duty at the individual level (e.g., in response to Singer 1972), we now suggest that institutional longtermism might mitigate them. So, if you are sceptical of individual longtermism because of the worries discussed below, here are some reasons why they might not apply to institutional longtermism.

First, some argue that a duty to do good sometimes falls short in collective action cases, where we can have a large impact as a collective but no (or little) impact as individuals. For example, consequentialists often argue we should go vegetarian because factory farming causes so much suffering. However, critics suggest that individual consumption decisions make no causal difference (Budolfson 2019; Schmidt 2024a; also see Barrett and Raskoff 2023). It is not implausible to think that such collective action problems can arise for individual longtermism, given the need for collective action to achieve many longtermist goals. But this challenge is less severe for institutional longtermism, since one function of institutions is to overcome such problems.[13]

Second, one consequentialist response to collective action worries is that even when individual actions seem to make no difference, they may have a tiny chance of making a huge difference. Refraining from buying chicken usually saves no one, but it may occasionally trigger a threshold in the supply chain that saves a huge number of chickens. If so, beneficence may demand vegetarianism because it does a lot of good *in expectation* (Singer 1980; Matheny 2002; Norcross 2004; Kagan 2011). However, for individual longtermism, this response to collective action problems seems to require a decision-theoretic commitment that is controversial in other contexts: *fanaticism* (see Greaves and MacAskill this volume: sec. 8 and references therein). Suppose you must choose between saving a hundred lives for sure and saving a quadrillion lives with a one-in-a-trillion probability (and assume saving lives has constant marginal value). According to expected value theory, you should choose the latter, since a one-in-a-trillion chance of saving a quadrillion lives is just as good as certainly saving a *thousand* lives. But this verdict is often seen as counterintuitively 'fanatical' in its pursuit of tiny probabilities of enormous value.

Now, predicting and controlling the long-run future is difficult, and many attempts to do so may have slim chances of success. For example, donating to an AI safety organization may only have a tiny chance of making a (huge) difference. Individual longtermism might therefore require fanaticism (see Tarsney 2023 for discussion). Of course, responses are available: one could defend fanaticism (Beckstead and Thomas 2023; Wilkinson 2022; see Russell 2024; Barrett 2024a for responses) or argue that, empirically, longtermist actions have a non-fanatical probability of doing long-run good. Regardless, institutional longtermism can often avoid such difficulties, as institutions typically have a much larger chance of making a difference than individuals (compare Kosonen 2023).

Third, consequentialism is often criticized for being too demanding: it does not leave room for personal pursuits, relationships, and projects (see Sobel 2020 for discussion). And even a somewhat stringent *pro tanto* duty of beneficence can raise similar worries. Now, if individual longtermism is implied by beneficence, this worry may intensify: the huge stakes associated with the long-term future may render beneficence even more demanding, and certain familiar strategies for avoiding demandingness problems may no longer be available (Mogensen 2021). However, institutional longtermism softens such demandingness problems, as there is less reason to worry about too much being demanded of *institutions*. Moreover, institutions might lighten the total burden individuals face by promoting

[13] Of course, collective action problems might still arise *between* institutions, for example, in getting various nations to do their part to combat climate change. Our claim is not that institutions resolve all collective action problems; it is only that they help.

longtermist aims more effectively, for example, due to economies of scale or a division of labour (Buchanan 1996; Goodin 2017).[14]

Fourth, many of our most meaningful projects and priorities involve relationships with others, such as our friends and family. Consequentialism or a weighty duty of impartial beneficence might unduly shrink the space for partiality. Of course, consequentialists respond to such worries (e.g., Railton 1984; Jackson 1991). But institutional longtermism can again sidestep them. Rather than to individuals, it applies impartial beneficence to institutions. And, as Robert Goodin argues, while being partial towards one's family and friends carries intuitive ethical weight for individuals, such partiality is objectionable at the institutional level. We expect private citizens to be partial but legislators to set aside personal relations and allegiances when designing or enacting law and policy (Goodin 1995; also see Pettit 2012).[15]

The above are sketches of arguments, not fully worked-out versions thereof. Still, further explorations of them seem worthwhile, since they suggest that the beneficence-based argument might be more robust for institutional than for individual longtermism.

3 Institutional longtermism meets mainstream political philosophy

So far, we have focused on the beneficence-based case for institutional longtermism. However, most political philosophers are more concerned with other considerations, like justice, legitimacy, or democracy. Might these defeat institutional longtermism?

3.1 The Stakes-Sensitivity Argument

Here, a standard longtermist argument (adapted from Greaves and MacAskill this volume) is the Stakes-Sensitivity Argument:

A. When the axiological stakes are very high, non-consequentialist considerations are (largely) outweighed, such that consequentialist reasons (largely) determine institutions' choiceworthiness.
B. When choosing among institutional options that significantly affect, in expectation, the long-term future, the axiological stakes are very high.
C. So, when choosing among institutional options that significantly affect, in expectation, the long-term future, consequentialist reasons (largely) determine institutions' choiceworthiness.

[14] One might still worry that longtermist institutions will demand more of individuals than non-longtermist ones. Whether this is so depends on complex issues that we cannot discuss here; for example, on some views, both sorts of institutions might instead be similarly demanding but simply demand different things. Regardless, our point is only the comparative one that demandingness worries are worse for individual than for institutional longtermism.

[15] Legislators must set aside *personal* relations, like friendships, but not necessarily other types of partiality, like being partial towards citizens of their own countries. Whether national partiality is justified and what it implies is an open question in general (Goodin 1988) and even more so under institutional longtermism.

Premise A allows for weighty non-consequentialist reasons of (say) justice and legitimacy but claims that they are overridden when the axiological stakes are high. This is intuitive. In emergency situations, when many lives are at stake—because of a war or a natural catastrophe, for example—it often seems justified to override 'non-consequentialist' concerns. Premise B adds that when institutions significantly affect the long-run future the stakes are indeed very high, since trillions or quadrillions of lives might be at stake in the long-run future.

While simple and powerful, many find the Stakes-Sensitivity Argument hard to accept. For example, Premise B is open to challenge from moral views on which large *numbers* of future people don't necessarily produce large *stakes*. Consider non-aggregative views on which 'the numbers don't count': we should be more concerned with a larger harm to one person than any number of smaller harms to others (Scanlon 1998). Or consider 'person-affecting' views in population ethics on which non-existing people don't count the same as existing people: it is neither good nor bad to bring new happy people into existence (Narveson 1973). Such views, and more sophisticated variants of them, are often thought to challenge longtermism.[16]

These debates are well-explored elsewhere. So, here we proceed on the assumption that the above objections are not decisive, and instead investigate how distinctively *political* values interact with longtermism. The Stakes-Sensitivity Argument suggests that even non-consequentialists cannot brush aside long-term consequentialist considerations so easily. But many political philosophers think that non-consequentialist considerations weigh very heavily. John Rawls, for example, famously identifies *justice* as the first virtue of institutions (Rawls 1971). If such considerations indeed weigh heavily, they might defeat longtermism (or at least its strong version on which promoting long-term value is *the* key priority). Alternatively, political philosophy may furnish new arguments in favour of longtermism, since many political values also call for the consideration of future people. For example, we might care about intergenerational justice or about the freedom and equality of future people. And how compelling is an ideal of democracy that disenfranchises most people, namely all those still to come?

In the remainder of this chapter, we therefore focus on how longtermism interacts with five mainstream values in political philosophy: justice, equality, freedom, legitimacy, and democracy. In each case, we first consider possible tensions with longtermism and then arguments pointing toward greater convergence.

3.2 Justice

T. M. Scanlon distinguishes between morality in general and *interpersonal morality* which is concerned with directed duties or *what we owe to each other* (Scanlon 1998: 6–7). Justice, according to Scanlon, falls within interpersonal morality, whereas at least some forms of impartial beneficence do not. Beneficence can involve undirected duties or reasons to

[16] See, e.g., Heikkinen (2022) and Curran (this volume) on non-aggregative views and Thomas (2023) on population ethics. Of course, it is also possible to challenge the relevance of the Stakes-Sensitivity Argument by appealing to empirical views on which institutions cannot significantly affect the long-run future.

promote value—duties not owed *by* anyone in particular, or *to* anyone in particular. But justice is distinctively concerned with what people are *due* (Gilabert 2016; Miller 2021; Barrett 2022).[17]

Already, this generates a tension between longtermism and justice. The beneficence-based case for institutional longtermism at least *prima facie* appeals to undirected reasons that are not owed to anyone in particular, as they pertain to possible people in the far future (though more on this shortly). Justice, in contrast, is owed to someone.

Besides this *structural mismatch* there might also be an *intuitive conflict* between justice and longtermism. Why should we worry so much about possible people in the far future when there are so many pressing injustices here and now? Shouldn't we address ongoing and historical injustice first?

However, several considerations might lessen these tensions.

First, longtermists might argue that longtermist priorities perform well in terms of near-term justice. Consider pandemic preparedness. As the COVID-19 pandemic shows, pandemics are supremely bad in the near term and lead to massive injustices, involving unnecessary death, illness, and poverty. And their burdens tend to fall disproportionately on disadvantaged individuals and groups. Similarly, AI safety work might also prevent near-term injustice (say, due to algorithmic discrimination). More generally, Carl Shulman and Elliott Thornley argue that most longtermist interventions that reduce extinction risks (or other similarly catastrophic risks) are competitive given the cost-benefit standards of rich countries, even if one only considers present people (Shulman and Thornley this volume).

Second, there might be institutional convergence: just institutions may better promote long-term value. For example, states that protect human rights, secure some decent economic minimum, and observe principles of legal justice might perform better by longtermists' lights. Furthermore, from a patient perspective, unjust institutions may be sticky and involve feedback loops that block institutional improvements (Barrett and Buchanan 2024). For example, those who benefit from unjust power inequalities tend to shape institutions in ways that further their short-term interests. And societies subject to epistemic injustice may be worse at institutional learning.

Besides these (speculative) empirical considerations, there are also more philosophical considerations pointing toward convergence.

First, justice gives us directed duties we owe to others. But, like beneficence, it may also give us undirected duties to *promote* justice. For example, imagine a button that would immediately remove injustices in some far-away country (or that would prevent them from occurring in the far future). Regardless of whether you have directed duties to people there, it seems you ought to press it. Rawls, too, thought that we have a 'natural duty' of justice to promote or sustain just institutions (Rawls 1971: 98–99).

If there is an undirected duty to promote justice, then presumably we have stronger reasons to promote more justice. Roger Crisp and Theron Pummer thus suggest that we should focus on reducing injustice in low-income countries, since, at the margin, we can

[17] Directed duties, as we use the term, differ from undirected duties in two ways. First, they concern what we owe *to* particular individuals, such that when we violate these duties we not only *do wrong* in an impersonal sense, but *wrong* someone in particular. On many views, duties of beneficence need not be directed in this way. Second, directed duties may also involve agent-relativity. For example, justice may require that *I* compensate you for my past wrongdoing, and not just that you are compensated by *someone*. Duties of beneficence, in contrast, are typically thought to be agent-neutral.

typically reduce injustice more effectively there than in high-income countries (Crisp and Pummer 2020: 401–402). However, if longtermists are right, our expected impact might be largest on future people, given their great number. An undirected duty to promote justice might push us to be 'justice longtermists'.

What exactly justice longtermism implies, however, is complicated. For example, some suggest that injustice is primary, with justice being merely the absence of injustice (e.g., Shklar 1990; Schmidtz 2011). If so, we can have a duty to prevent injustice, but no duty to promote justice. Does justice longtermism then imply a duty to hasten human extinction, since without people, there can be no injustice? This may strike many as a *reductio*.[18] Alternatively, we might have a duty to promote justice rather than to prevent its opposite. But does that imply we ought to bring more people into existence, to increase aggregate justice among them? We are not used to thinking of larger countries as more just simply in virtue of their greater population. Clearly, some work is needed on the 'population ethics of justice'.

A second type of convergence argument zooms in on directed duties of *intergenerational justice*: duties we owe to future people. There is some debate about whether we can owe future people anything in light of the non-identity problem: what we do now affects not only how well off future people are, but also *which* future people exist (see Meyer 2021: sec. 3 for an overview). But many political philosophers believe that we at least owe it to future people not to bring them into existence beneath some threshold of sufficiency or in circumstances where their rights will be violated (see Caney 2018 for discussion). Moreover, conditional on there being future people, we might have directed duties towards them, if we think of these duties as owed to 'types of person' or persons *de dicto* (Hare 2007; Kumar 2015; 2018). However, it is harder to see how we can have directed duties of justice *to* future people to bring them into existence in the first place. To *whom* have we acted unjustly if none come to exist? The tension between intergenerational justice and the longtermist priority of avoiding human extinction may therefore persist (Barrett 2022).[19]

Another issue is that many global justice theorists argue that (certain) duties of distributive justice are owed only to those we bear special relations to. For example, Michael Blake and Laura Valentini argue that we only have duties of (egalitarian) justice towards others if we share *coercive* institutions with them (Blake 2001; 2013; Valentini 2011a; 2011b). Andrea Sangiovanni argues that the relevant relation is *reciprocity* (Sangiovanni 2007). Can such relations exist across generations? Intuitively, relations of direct reciprocity cannot (Heyd 2009), but recent work suggests that relations of indirect reciprocity might (Gosseries 2009; Heath 2013; Brandstedt 2015; Scheffler 2018). Yet even if such relations extend across generations, they are unlikely to get us all the way to (full) longtermism. Intergenerational coercion, reciprocity, and so on are likely weaker across than within generations, and may weaken and even disappear as we peer further into the future (compare Mogensen 2022).

Finally, even if duties of intergenerational justice extend into the far future, they might require something different than our undirected duties to promote valuable or just outcomes.

[18] Even if justice longtermism implies a duty to hasten human extinction, this duty would only be *pro tanto*, so it wouldn't imply that we all-things-considered ought to do this. Still, we suspect that many will find even this qualified implication hard to stomach.

[19] While failing to prevent human extinction might also be unjust to *present* humans, this convergence may not support (for example) the longtermist thought that even reducing the probability of human extinction by a small probability is hugely important, or that preventing outright extinction may be much more important than preventing near-extinction events from which we could recover.

Many theories of intergenerational justice are sufficientarian, only requiring us to ensure that future people meet some minimal standard (Meyer and Roser 2009). Undirected duties to promote justice and the good within future generations might not be limited this way. Ultimately, this depends on our particular theory of justice. Although we cannot provide a survey of such theories here, we now explore the two most common values used to fill out theories of justice, but which may also matter for other reasons: equality and freedom.

3.3 Equality

Equality is a central value in political philosophy, perhaps because there are so many reasons to care about it (Miller 1997; O'Neill 2008; Scanlon 2018). We here focus on three egalitarian views, which are distinct though not mutually exclusive.

First, a more equal distribution of income and wealth might be instrumentally valuable by contributing to more well-being, trust, education, better political institutions, social mobility, and the like (Woodard 2019: ch. 7; Schmidt and Juijn 2024). Call this *instrumental egalitarianism*.

Second, *relational egalitarianism* holds that what matters is establishing equal relations and preventing problematic relational inequalities.[20] Relations like domination and subjugation are the central enemies, while securing conditions for people to live as equals is the positive ideal. For relational egalitarians, distributive inequalities are not bad in themselves, but they are objectionable if they constitute or contribute to relational inequalities.

Third, distributive egalitarians hold that distributive inequalities are non-instrumentally bad.[21] Luck egalitarians add that inequalities between individuals are only bad if they arise from brute luck but not if they arise from responsible choice.[22]

How do these versions of egalitarianism bear on longtermism? To answer this, we must consider, (i) whether each view applies only within or also across generations; and (ii) what kind of reasons each view gives us: undirected reasons to promote good outcomes or directed reasons owed to others?

First, answering (i), some *instrumental* egalitarian arguments apply to both intragenerational and intergenerational distributions. For example, more equal distributions of economic resources between generations might lead to more well-being because of such resources' decreasing marginal utility. So, if we expect future people to be richer than us, this might push us to spend somewhat more on present people's welfare. Other arguments apply more to intragenerational distributions, such as arguments around relative standing and status anxiety. On (ii), instrumental egalitarianism is about good outcomes and not about directed duties of justice. So there is no obvious structural tension between instrumental egalitarianism and longtermism.

However, Tyler Cowen argues that for longtermists, economic (in)equality is mostly irrelevant (Cowen 2018): sustainable growth is far more important, as it compounds

[20] See, e.g., Young (1990), Anderson (1999), Fourie, Schuppert, and Wallimann-Helmer (2015), Lippert-Rasmussen (2018), Schemmel (2021), Schmidt (2022c).

[21] Alternatively, distributive egalitarians might hold that distributive equalities are non-instrumentally good. This raises issues similar to those discussed under 'justice longtermism': if inequality is bad, we might have reason to hasten human extinction. See Arrhenius and Mosquera (2022) and Mogensen (this volume).

[22] See, e.g., Arneson (1989), Cohen (1989), Stemplowska (2013), Lippert-Rasmussen (2015).

over the years. Andreas T. Schmidt and Daan Juijn disagree, arguing that reducing (intragenerational) economic inequality is probably valuable (even assuming a utilitarian axiology) whether one takes a short, medium, or longtermist time frame (Schmidt and Juijn 2024). Central arguments here are that more equal societies likely have lower greenhouse gas emissions, lower risk of elite capture of political institutions, and better all-purpose conditions for public good provision and for dealing with long-term risks.

Consider *relational* egalitarianism next. Regarding (i), most relational egalitarians focus on intragenerational inequalities, but some relational inequalities might obtain across generations too (e.g., Bengtson 2019). Regarding (ii), relational egalitarianism is sometimes considered a theory of interpersonal justice and sometimes an axiological theory. Insofar as relational egalitarianism is interpersonal and about directed duties, it might conflict with longtermism: the strongest relationships will likely be intragenerational, and stronger intergenerational relationships will be near-term rather than long-term. However, if relational egalitarianism issues undirected duties to promote justice, it might not clash with longtermism but merely affect its shape: preventing relational inequalities among future people would become a longtermist priority.

Finally, consider *distributive* egalitarianism. Regarding (i), the theory can be both intra- and intergenerational, though a concern with intergenerational equality sometimes leads to counterintuitive verdicts (Temkin 1995: 99; Schmidt 2024b). Derek Parfit, for example, asks whether it really matters that 13th-century Inca peasants were worse off than people alive today (Parfit 1991: 7; though see Segall 2016). Regarding (ii), distributive egalitarianism is typically understood to be about the value of outcomes. So there is no structural conflict with longtermism, and distributive egalitarianism might only affect longtermism's shape.[23]

3.4 Freedom

Freedom is another central value in contemporary political philosophy. On first glance, it seems to conflict with longtermism: freedom is often thought to place limits on using state power to bring about certain outcomes. Longtermism, however, might require that we infringe the freedom of existing people.

Nick Bostrom provides an extreme example of a potential clash (Bostrom 2019). He asks whether we may have been lucky so far to have only discovered ways of causing global disasters that are costly and difficult to implement. But what if the nuclear weapons or pandemic-causing pathogens of the future could be built in your garage? To reduce the risk that such technological developments occur, Bostrom wonders whether we might have to violate people's freedom and privacy on a massive scale, for example, through a global surveillance state. Beyond such horrifying scenarios, there are more mundane cases where longtermism might conflict with freedom. For example, might preventing long-term damage from climate change require restricting people's freedom now?

[23] If distributive egalitarians give *overriding* weight to equality, or perhaps to benefiting the worst off, this might challenge longtermism rather than just change its shape, at least assuming future people will be better off than us. Arguably, however, all plausible versions of distributive egalitarianism allow that large increases in aggregate welfare (as are at stake in the very long run) sometimes outweigh inequality or the interests of the worst off (Barrett 2020a; 2020b).

To explore these issues, we need a better handle on what freedom is and why we should care about it. Contemporary political philosophy often distinguishes between liberal, republican, and libertarian freedom (Schmidt 2022d). Liberal theorists hold that freedom is primarily about having *options*. Some liberals hold that the absence of interpersonal interference with your options is sufficient for freedom (Miller 1983; Steiner 1994; Kristjánsson 1996). Others hold that freedom requires not only the absence of interference, but also the genuine ability to pursue an option (Parijs 1997; Sen 1999; Kramer 2003; Schmidt 2016). Recent republican theories argue that even this is insufficient: freedom also depends on not being subject to domination, where someone dominates you if they have the uncontrolled power to interfere with you—regardless of whether they actually intervene (Lovett 2010; Skinner 2012; Pettit 2014; Schmidt 2018a).

Practically, liberal and republican theories of freedom mainly converge, first, because (most) republicans think option-freedom is necessary (but insufficient) and, second, because most liberals think republican institutions and relations of non-domination tend to increase liberal option-freedom (Carter 1999: ch. 7.5; Kramer 2003: ch. 3). Interestingly, recent theorizing in this tradition often treats freedom as a scalar good rather than something imposing a deontic constraint (Schmidt 2022d).[24] If we go with liberal or republican theories, then, longtermism and the pursuit of freedom might seem to converge: reasons to care about the freedom of existing people are also reasons to care about future people's freedom, whether this is understood as option-freedom or non-domination (Vercelli 1998; Schmidt 2025). Some republicans add that there can also be domination *across* generations (Smith 2013; Beckman 2016; Katz 2017; Schmidt and Bengtson 2021). Previous generations dominate future people, as they have the uncontrolled power to influence their lives. All this suggests that freedom might not speak against longtermism but only affect its shape.

But perhaps we are assuming an overly consequentialist view on freedom. Rather than promoting other people's freedom, aren't we required to respect it?

There are different deontological theories on offer but the most influential are libertarian. Libertarians see freedom as intrinsically linked to *property rights*: we are unfree to the extent that our property rights are violated, including our right of self-ownership (Nozick 1974; Vallentyne and Steiner 2000; Otsuka 2003; Fried 2004). If such views issue deontological constraints, they might limit the scope within which institutions can permissibly pursue longtermist causes. But even this is not clear.

First, libertarians often allow interfering with person *A*'s actions if this interference is necessary to prevent interference with person *B*'s freedom. For example, if *A* threatens *B*'s physical safety, libertarians might endorse restricting *A*'s options to safeguard *B*'s freedom. So perhaps some restrictions now might be necessary to reduce the risk of restrictions on future people's freedom. Second, even if we see freedom as issuing deontological directed duties, it might also give us undirected duties to promote it. Shouldn't libertarians try to bring about libertarian institutions for future people? Furthermore, most libertarians are not entirely insensitive to axiological stakes. Even Robert Nozick held that preventing 'catastrophic moral horrors' might override freedom and property rights (Nozick 1974: 29).

[24] As we did for justice and equality, we might ask: is freedom good, or unfreedom bad? Each answer—if it invokes non-instrumental value or disvalue—will encounter the by now familiar complications in variable population cases.

Longtermists could argue that they are concerned with preventing such horrors from occurring, or with promoting outcomes that it would be equally disastrous not to achieve.

Leaving philosophical considerations aside, there may also be strong empirical reasons why longtermists should safeguard freedom (of both current and future people) (see Schmidt 2025, pp. 205–6). First, empirical evidence suggests societies with more freedom are, on the whole, happier (Veenhoven 2000; Inglehart et al. 2008; Bavetta et al. 2014). Moreover, most rich countries (other than tax havens and petrol states) are broadly speaking liberal democracies. This suggests freedom has benefits at least in the medium term and perhaps in the very long run. Second, the risk of 'value lock-in' might give longtermists reason to prefer freedom as a default. If we are uncertain about what the correct values are, we should be cautious not to lock in values for a long time that later might turn out misguided. Free societies are less likely to lock in contested values and may offer better epistemic conditions for experimentation and moral progress. Finally, and relatedly, longtermists worry about totalitarian risk. A strong societal and institutional commitment to freedom might reduce such risk.

3.5 Legitimacy

So far, we have seen that while justice, equality, and freedom pose *prima facie* challenges to longtermism, various arguments also suggest convergence. However, for legitimacy, the tension with longtermism seems stark. On common theories of legitimacy, a state is legitimate only if it rules by the consent of the governed (e.g., Locke 1690/1990; Simmons 2001), or if its actions are the output of a fair democratic procedure (e.g., Christiano 2008), or if its constitution or laws are 'publicly justified'—that is, justified *to* all reasonable people it rules over, in light of their diverse values and beliefs (e.g., Rawls 2005). Many actions favoured by longtermists might therefore prove illegitimate in modern societies, where citizens do not generally consent to, vote for, or agree with longtermist priorities.

Moreover, the *structural* tension with longtermism also seems greater for legitimacy: unlike the values discussed above, legitimacy is uncontroversially deontological. A government or other entity is legitimate when it has the 'right to rule'. Minimally, this is a 'permission-right' implying it permissibly wields its political power. More controversially, legitimacy may involve *authority* or a 'claim-right' that implies individuals have an obligation to obey. Although we might also have some reason to promote legitimacy, legitimacy's force is widely assumed to be deontological, so repeating the above move of making it an object to promote seems less promising.

Nevertheless, we might try to resolve this tension.

First, we might argue that, empirically, pursuing certain longtermist priorities is indeed legitimate, at least on some theories. For example, perhaps efforts to combat extinction risks are publicly justified, since they also severely threaten present people. Less speculatively, Eric Martinez and Christoph Winter present survey evidence that most legal professionals and laypeople believe the law both can and should protect future people much more than it currently does (Martinez and Winter this volume). This tentatively suggests a case for the legitimacy of relevant legal reforms.

Second, and more philosophically, we might offer a modified version of the Stakes-Sensitivity Argument, where stakes-sensitivity is built into the very concept of legitimacy

(rather than, as on the standard argument, stakes potentially outweighing legitimacy). Ross Mittiga argues that, when the stakes are high enough, ordinary notions of legitimacy allow the state to employ emergency powers that would be illegitimate in normal times (as occurred during the COVID-19 pandemic, for example) (Mittiga 2022). On his view, climate change may represent an ongoing state of emergency such that, in tackling it, governments can legitimately act in ways normally seen as illegitimate or even authoritarian. Longtermists might generalize this argument, claiming that the same applies to efforts to promote massive long-run value. However, most theorists would likely deny that legitimacy really is stakes-sensitive in this way: while many agree that states can render themselves *illegitimate* by acting very badly or unjustly (e.g., Rawls 2005: 428), few agree that they can legitimate themselves by doing enough good. Furthermore, as seen above, societies that value freedom are likely to promote long-term value better than those that do not. And as we will see, the same may be true of democratic rather than authoritarian states. Indeed, a government that effectively declared a perpetual state of emergency should raise red flags to longtermists due to the risk of bad value lock-in, long-term totalitarian capture, and other troubling path-dependencies.

Finally, and most radically, we might challenge existing approaches to legitimacy as being too focused on present people. Allen Buchanan, for example, ties legitimacy to justice, defending the 'functionalist' thesis that a state is legitimate when it meets minimal standards of justice. One of his central arguments is that the state is such a dangerous and powerful entity that only justice is weighty enough to legitimate its establishment (Buchanan 2003: 247). Longtermists may retort that the value at stake in the long-run future is also of great weight, such that similar arguments should lead one to view a state as legitimate only if it appropriately pursues long-run value.

Tyler John suggests another argument pointing in the same direction. He notes that the demand for legitimacy arises given some relation between the ruler and the ruled, and argues that whatever the relevant relation is in the intragenerational case—be it coercion, domination, or subordination, say—the same relation holds intergenerationally (John 2022: ch. 1). Others demur. Ludvig Beckman, for example, argues that present governments cannot 'rule' future people at all, because future people can repeal or modify laws we pass now (Beckman 2009: ch. 7). Furthermore, it is not obvious how states *can* legitimate themselves to future people, at least on certain theories. For example, future people cannot consent to what present states do (although perhaps we can appeal to notions like hypothetical consent). Still, if states must indeed legitimate themselves to future people, then legitimacy arguably provides an argument *for* longtermism. Consider democratic conceptions on which states derive legitimacy from the participation or representation of its subjects in appropriate decision-making procedures. If future people need to be represented too, then this might radically change which actions are legitimate, perhaps suggesting that longtermist institutions and policies are not only consistent with, but mandated by, legitimacy. This leads into our next topic.

3.6 Democracy

Intuitively, respecting the voice and preferences of existing people might constrain how longtermist institutions can be. Stephen A. Marglin, for example, appealed to democracy when defending a positive social discount rate, on the grounds that existing people discount

the future (Marglin 1963). Yet, again, it might also seem undemocratic that the vast number of future people are disenfranchised. So, does democracy push us towards or away from longtermism?

The answer partly depends on *why* one supports democracy. Democracy is defended on two main sorts of grounds: intrinsic and instrumentalist (though see Ziliotti (2020) for more fine-grained distinctions). Intrinsic defences point to features of democratic decision-making procedures that are valuable independently of the outcomes they produce. For example, such procedures may realize values of freedom, equality, collective self-rule, or fairness. Instrumental defences appeal to the beneficial consequences of democracy. For example, democracies may have good empirical track records or have epistemic properties that yield better decisions than other systems.

In practice, 'pure proceduralists' are rare. Most endorse both intrinsic and instrumental arguments. John Halstead thus argues that nearly all major theories of democracy are committed to 'high-stakes instrumentalism': when the moral stakes are very high, the intrinsic value of democracy is outweighed, and we ought to employ whatever political procedures yield better outcomes (Halstead 2017). Given that longtermist decisions involve high stakes, the Stakes-Sensitivity Argument therefore threatens to collapse all democratic theories to instrumentalism.

Suppose, however, that this collapse can be avoided. Do intrinsic theories of democracy then conflict with longtermism? On the one hand, the conflict between democracy and longtermism seems obvious: democratic bodies may predictably decide against longtermist priorities. On the other, actual democracies might fall short of the intrinsic ideals of democracy. Assume democracy implies, roughly, that all constituents enjoy equal say or representation in decision-making. The question now arises *who* should be included in this constituency. The two most prominent approaches to answering this 'boundary problem' involve the 'all-affected' and the 'all-subjected' principles. According to the former, all individuals whose interests are (actually or possibly) affected by a decision should have influence over or representation in the decision (Arrhenius 2005; Goodin 2007). According the latter, all who are bound by, subject to, or coerced by a decision should be included (Beckman 2014: 257; Goodin 2016: 370–373).

Do these principles imply we must include future people? Initially, it seems so: future people seem both affected by and subject to current decisions. However, the non-identity problem may suggest otherwise for the all-affected principle (Tännsjö 2007). And Beckman's above argument that we cannot rule over future people threatens the application of the all-subjected principle (Beckman 2009: ch. 7). Further, these two principles are both highly revisionary and somewhat 'free-floating', and several theorists now argue that we should look beyond them and adopt more theory-driven approaches to the boundary problem that connect to a theory of what democracy is and why it is valuable (Saunders 2012; Song 2012; Miller 2020; Bengtson and Lippert-Rasmussen 2021). For example, some recent work explores whether and how relational egalitarian and republican theories of democracy should include future people (Schmidt and Bengtson 2021; Bengtson forthcoming).

Turn now to *instrumental* defences of democracy. In general, longtermists should care about at least some ways democracy is instrumentally valuable. For example, democracies arguably facilitate economic growth (Acemoglu et al. 2018), have better human rights records (Herre and Roser 2013), and correlate with better education, health (Herre and Roser

2013; Ortiz-Ospina 2019), happiness, and life satisfaction (Owen, Videras, and Willemsen 2008; Orviska, Caplanova, and Hudson 2014). Insofar as those benefits are real and democracies are sticky, democracies may provide steady streams of such benefits over time. Also pertinent to longtermism, 'democratic peace theory' holds that democracies go to war less often, especially with each other (Chan 1997). They might also better deal with certain disasters (see Sen (1982) on democracies and famines and Rubin (2009) for a response). Some also argue that democracies improve and adapt faster since they are better at experimenting and gathering feedback (Dewey 1927; Anderson 2006; Knight and Johnson 2011); others argue that they are better at resisting elite capture (Bagg 2018; 2024).

But are democracies really equipped to handle longtermist problems, including low-probability/high-impact risks that require long-term strategy? Electoral incentives notoriously focus politicians on the next election cycle, and mechanisms of democratic accountability generally hold politicians to account to present citizens rather than to future people (Caney 2017; John and MacAskill 2021). Some thus argue for more centralized or even authoritarian forms of governance to respond to the climate crisis (Mittiga 2022; see Shahar 2015 for an overview and critique of such 'eco-authoritarianism'). However, as defenders of this view themselves often acknowledge, authoritarian governments have poor track records on environmental issues (Shahar 2015: 354–356; Mittiga 2022: fns. 1, 2). Furthermore, the only systematic attempt we are aware of to quantify how well governments pursue long-run goals finds a strong *positive* correlation between democracy and 'Intergenerational Solidarity' (Krznaric 2021: ch. 9). If we also consider the above instrumental arguments plus worries around 'totalitarian risk' and value lock-in, longtermism seems more likely to reinforce rather than challenge the instrumental case for democracy.

Finally, longtermism raises questions for so-called 'epistemic' defences of democracy, including those that invoke formal results rather than empirical evidence (Cohen, 1986; Anderson 2006; Estlund 2009). The most famous defence rests on the Condorcet Jury Theorem, which says that under certain conditions the majority is more likely to be right than any individual (List and Goodin 2001; Goodin and Spiekermann 2018). Others have invoked the Diversity Trumps Ability theorem (Landemore 2012), which says that under certain conditions more diverse groups outperform groups composed of experts (Hong and Page 2004; Page 2008). Both theorems have their limitations, but the basic idea—that democracy can harness diversity and the wisdom of crowds—has staying power. Might longtermism challenge this? Speaking roughly, the key insight behind both theorems is that decision-making bodies face a trade-off between the greater competence (or expertise) of their members, and their greater diversity.[25] Crucially, however, the benefits of diversity only kick in if all members are 'competent enough'. Now, figuring out how to promote long-run value is very hard. This might seem to undermine the 'competent enough' condition and thus suggest an argument for epistocracy ('rule by the knowers'): if lay people are incompetent, we must leave longtermist governance to experts rather than the people. Conversely, however, longtermist problems might be tractable enough that most individuals do count as competent. Instead, the real challenge might be that no one is *very* competent to pursue

[25] However, the theorems do interpret diversity differently. The Condorcet Jury Theorem suggests that diversity helps insofar as it facilitates probabilistically independent judgements, whose errors can 'cancel out'. The Diversity Trumps Ability theorem operates instead through a 'baton-passing' mechanism that relies on a diversity in perspectives and heuristics: more diverse groups are less likely to get stuck at suboptimal solutions, since there is more likely to be at least one person who can find an improvement.

them: even 'experts' aren't much better than laypeople. If so, this might reinforce epistemic arguments for democracy, since it suggests we must rely on diverse crowds rather than (only on) expertise (see Ahissar 2022).

4 What next?

We have explored not only the beneficence-based argument for institutional longtermism, but also various points of tension and convergence between institutional longtermism and central values in political philosophy. In each case, we have found that while certain tensions initially seem manifest, on closer inspection things quickly get complicated, leaving it far from obvious whether political values conflict with, or even support, institutional longtermism.

Of necessity, our survey of these topics has been superficial as well as incomplete. Further research is needed, including more careful investigations both of the arguments we have floated and of what longtermism might imply for particular theories of justice, legitimacy, and the like.[26] Such work might also more carefully distinguish strong from weak (and full from medium) versions of longtermism, which we have often run together here. But rather than diving deeper into these topics, we end with a grab bag of other important questions in longtermist political philosophy, starting with more theoretical and broader questions:

Global justice and global governance: as noted, global justice and longtermism intersect in interesting ways. But longtermism also raises fascinating questions about global *governance*. Effective longtermist action—on issues like climate change, pandemic prevention, AI safety, and more—likely requires international action (Ord 2020). However, some worry that centralizing political authority raises totalitarian risk (Caplan 2011). So, should longtermists favour more or less international coordination or centralization, and of what sort?[27]

Political morality: we have focused on institutions rather than on how individuals should act. But how does longtermism affect political morality? For example, if longtermism holds up yet states fail to meet their longtermist duties, what does that imply for citizens and their political obligations and potential civic duties to effect change?

Diversifying: we have focused on contemporary analytic philosophy. But we might also gain insights from other periods and traditions. Consider first examples from the history of Western philosophy. John Stuart Mill (Mill 1866: cols 1525–1528) argued that we should leave coal in the ground for future generations—an argument recently explored by MacAskill (MacAskill 2022: 138–141). Edmund Burke (1790) provided a conservative political argument for concern for the future, which has been taken up by Ord (2020: 49–52). Hans Jonas developed a Kantian variant of longtermism several decades ago (Jonas 1979; 1985).

[26] For example, Porter and Gibbons (2024) consider how an appreciation of the longtermist priority of mitigating extinction (or other catastrophic) risks might lead parties behind a Rawlsian veil of ignorance to endorse different principles of justice than Rawls himself derived.

[27] Such questions are pressing: marking its 75th anniversary, the United Nations released a report that explicitly includes longtermist goals as central to the UN's mission (United Nations 2021). With more meetings and public deliberations planned, the UN will also consider concrete proposals to represent future generations, including a Trusteeship Council, a Futures Lab, a Declaration on Future Generations, and a Special Envoy.

Beyond the 'Western' tradition, the oral constitution of the Iriquous confederacy (*Gayanashagowa*) dates back centuries and includes concern for future generations, often interpreted as the 'seventh generation principle'. There is also much to learn by investigating points of contact between longtermism and other traditions, such as Buddhist ethics (Baker 2022: sec. 3.2.2), Confucianism (Hourdequin and Wong 2021), Latin American thought (Vidiella and García Valverde 2021), African thought (Mbonda and Ngosso 2021), and Thomist Christian thought (Riedener 2022).

Non-human animals: non-human animals remain neglected by political philosophers (Donaldson and Kymlicka 2011; Garner 2013; Schmidt 2018b; Barrett 2022) and longtermists (Browning and Veit this volume). This is unfortunate, since serious efforts to promote longtermist institutional reform must take animals into account. It also raises challenging questions (MacAskill 2022: 208–213): given humanity's impact on non-human animals and ecosystems, is the value of our survival really positive on balance? Furthermore, bringing non-human animals and longtermism together might uncover sources of insight, since the political issues confronting non-human animals and future people have much in common: despite their massive numbers, both groups are utterly disenfranchised.

Distinctively longtermist values: we have discussed how longtermism interacts with existing values in mainstream political philosophy. But might longtermism also require new, specifically longtermist values to shape our institutions? If so, what might those be? For example, should we focus more on adaptability?

Longtermist institutional reform: what longtermist political reforms, if any, should we pursue (González-Ricoy and Gosseries 2017; John and MacAskill 2021)? For example, should we focus on *constitutions* (Beckman 2016; Araújo and Koessler 2021) or more on the legislative or executive? Or should we focus less on formal changes, and more on informal norms, or on having civil society and voters pressure decision-makers into longtermist action?

There are also important questions about concrete longtermist priorities:

Pandemics and public health ethics: public health ethicists have already written much about the ethics and politics of infectious diseases and pandemics (Boylan 2022; Hirose 2022). Integration and further work might be important for targeted interventions to reduce pandemic risks that threaten long-run value.

Nuclear weapons: in the heyday of the Cold War, several philosophers wrote about nuclear weapons (Goodin 1980; Cohen and Lee 1986; Lewis 1986; Kavka 1987). Renewed interest—with a view towards longtermism and extinction risks—might be valuable, particularly in evaluating specific interventions (e.g., Rendall 2021).

Population ethics and demography: population ethics throws up challenging questions that receive much attention in the literature on longtermism. But applied political questions around demography also demand our attention. Near-term worries tend to cluster around overpopulation. For example, given the climate crisis, political philosophers discuss how many children one can permissibly have and what, if anything, states should do about

overpopulation (e.g., Conly 2016). However, if current trends continue, demographers predict that *depopulation* will replace overpopulation in the next century (Bricker and Ibbitson 2019; Jones 2022). So, what would a longtermist perspective imply here?

AI governance: for longtermists, AI presents both vast opportunities and risks. AI governance should thus be an important area for targeted interventions, and political philosophy could make valuable contributions to this growing field (e.g., Bullock et al. 2024).

Space governance: some longtermists view space settlement as a crucial step in humanity's future, since it may allow humanity to massively expand its population and fortify itself against certain catastrophic risks. The field of space governance, however, is in its infancy. Political philosophers could help.

Finally, consider some questions about *how* longtermism is pursued and promulgated, or about what we might call the politics (rather than political philosophy) of longtermism:

Longtermism as ideology: some worry that longtermism could come to serve as an ideology that, in the name of a long and glorious future, justifies maintaining or even worsening problematic features of the status quo. Notably, this criticism does not assume that longtermism is false, only that it might be abused (either intentionally or due to bias or motivated reasoning). How should we best understand, and guard against, this concern?

Democracy and longtermist philanthropy: currently, much investment into longtermism comes from private donors, including billionaires associated with effective altruism. Some worry that such funding structures are undemocratic and elitist (Reich, 2018; Lechterman 2021; Saunders-Hastings 2022; though see Barrett 2024b). Such worries are especially pressing in light of the recent collapse of and (alleged) fraud surrounding the FTX cryptocurrency exchange, given its longtermist philanthropic arm.[28]

This list remains only a sampling of the many questions and research avenues in longtermist political philosophy. Our goal has been to show that there is important work to be done here—work that we hope both longtermists and their critics will feel compelled to pursue. After all, the stakes may be very high.[29]

References

Abad, L. A. and Maurer, N. (2021), 'History Never Really Says Goodbye: A Critical Review of the Persistence Literature', in *Journal of Historical Political Economy* 1: 31–68.

Acemoglu, D. and Robinson, J. A. (2012), *Why Nations Fail: The Origins of Power, Prosperity, and Poverty* (Crown Business).

[28] However, if democratic worries primarily concern individual philanthropists circumventing the democratic process to pursue longtermist goals, then such worries may arise more for individual than for institutional longtermism.

[29] For helpful comments and discussion, we would like to thank Andreas Mogensen and Luchino Hagemeier. Thanks also to participants in the Global Priorities Institute Work in Progress Group and in the Centre for the Study of Social Justice Seminar Series, both at the University of Oxford.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2018), 'Democracy Does Cause Growth', in *Journal of Political Economy* 127: 47–100.
Ahissar, S. (2022), 'The Wisdom of the Crowds and the Long Term' (unpublished manuscript).
Anderson, E. (1999), 'What Is the Point of Equality?', in *Ethics* 109: 287–337.
Anderson, E. (2006), 'The Epistemology of Democracy', in *Episteme* 3: 8–22.
Araújo, R. and Koessler, L. (2021), 'The Rise of the Constitutional Protection of Future Generations', in *Legal Priorities Project Working Paper Series* 7: 1–44.
Arneson, R. J. (1989), 'Equality and Equal Opportunity for Welfare', in *Philosophical Studies* 56: 77–93.
Arrhenius, G. (2005), 'The Boundary Problem in Democratic Theory', in G. Arrhenius and F. Tersman (eds.), *Democracy Unbound: Basic Explorations* (Stockholm University), 14–29.
Arrhenius, G. and Mosquera, J. (2022), 'Positive Egalitarianism Reconsidered', in *Utilitas* 34: 19–38.
Bagg, S. (2018), 'The Power of the Multitude: Answering Epistemic Challenges to Democracy', in *American Political Science Review* 112: 891–904.
Bagg, S. (2024), *The Dispersion of Power: A Critical Realist Theory of Democracy* (Oxford University Press).
Baker, C. (2023), 'Buddhism and Utilitarianism', in R. Y. Chappell, D. Meissner, and W. MacAskill (eds.), *An Introduction to Utilitarianism*, https://www.utilitarianism.net/guest-essays/buddhism-and-utilitarianism (accessed March 2023).
Barrett, J. (2020a), 'Efficient Inequalities', in *Journal of Political Philosophy* 28: 181–198.
Barrett, J. (2020b), 'Is Maximin Egalitarian?', in *Synthese* 197: 817–837.
Barrett, J. (2020c), 'Social Reform in a Complex World', in *Journal of Ethics and Social Philosophy* 17: 103–132.
Barrett, J. (2022), 'Social Beneficence', GPI Working Paper 11-2022 (Global Priorities Institute, Oxford University).
Barrett, J. (2024a), 'In Defense of Moderation', GPI Working Paper 32-2024 (Global Priorities Institute, Oxford University). https://globalprioritiesinstitute.org/wp-content/uploads/Jacob-Barrett-In-Defence-of-Moderation.pdf (accessed March 2023).
Barrett, J. (2024b), 'Philanthropy for the Disenfranchised' (unpublished manuscript).
Barrett, J., and Buchanan, A. (2024), 'Social Experimentation in an Unjust World', in *Oxford Studies in Political Philosophy* 9: 127–152.
Barrett, J., and Gaus, G. F. (2020), 'Laws, Norms, and Public Justification: The Limits of Law as an Instrument of Reform', in S. A. Langvatn, M. Kumm, and W. Sadurski (eds.), *Public Reason and Courts* (Cambridge University Press), 201–228.
Barrett, J., and Raskoff, S. (2023), 'Ethical Veganism and Free Riding', in *Journal of Ethics and Social Philosophy* 24: 184–212.
Barry, B. (1997), 'Sustainability and Intergenerational Justice', in *Theoria* 44: 43–64.
Bavetta, S., Navarra, P., Maimone, D., and Patti, D. (2014), *Freedom and the Pursuit of Happiness: An Economic and Political Perspective* (Cambridge University Press).
Beckman, L. (2009), *The Frontiers of Democracy: The Right to Vote and its Limits* (Palgrave Macmillan).
Beckman, L. (2014), 'The Subjects of Collectively Binding Decisions: Democratic Inclusion and Extraterritorial Law', in *Ratio Juris* 27: 252–270.
Beckman, L. (2016), 'Power and Future People's Freedom: Intergenerational Domination, Climate Change, and Constitutionalism', in *Journal of Political Power* 9: 289–307.
Beckstead, N. and Thomas, T. (2023), 'A Paradox for Tiny Probabilities and Enormous Values', in Nous 58: 431–455.
Benatar, D. (2008), *Better Never to Have Been: The Harm of Coming Into Existence* (Oxford University Press).
Bengtson, A. (2019), 'On the Possibility (and Acceptability) of Paternalism towards Future People', in *Ethical Theory and Moral Practice* 22: 13–25.
Bengtson, A. (forthcoming), 'Finding a Fundamental Principle of Democratic Inclusion: Related, Not Affected or Subjected', in *Inquiry*.
Bengtson, A. and Lippert-Rasmussen, K. (2021), 'Why the All-Affected Principle Is Groundless', in *Journal of Moral Philosophy* 18: 571–596.
Blake, M. (2001), 'Distributive Justice, State Coercion, and Autonomy', in *Philosophy & Public Affairs* 30: 257–296.
Blake, M. (2013), *Justice and Foreign Policy* (Oxford University Press).
Bostrom, N. (2019), 'The Vulnerable World Hypothesis', in *Global Policy* 10: 455–476.
Boylan, M. (2022), *Ethical Public Health Policy within Pandemics: Theory and Practice in Ethical Pandemic Administration* (Springer Nature).
Brandstedt, E. (2015), 'The Circumstances of Intergenerational Justice', in *Moral Philosophy and Politics* 2: 33–55.

Bricker, D. and Ibbitson, J. (2019), *Empty Planet: The Shock of Global Population Decline* (McClelland & Stewart).
Broome, J. (1994), 'Discounting the Future', in *Philosophy & Public Affairs* 23: 128–156.
Browning, H. and Veit, W. (this volume), 'Longtermism and Animals', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Buchanan, A. (1996), 'Perfecting Imperfect Duties: Collective Action to Create Moral Obligations', in *Business Ethics Quarterly* 6: 27–42.
Buchanan, A. (2003), *Justice, Legitimacy, and Self-Determination: Moral Foundations for International Law* (Oxford University Press).
Budolfson, M. (2019), 'The Inefficacy Objection to Consequentialism and the Problem with the Expected Consequences Response', in *Philosophical Studies* 176: 1711–1724.
Bullock, J., Chen, Y., Himmelreich, J., Hudson, V. M., Korinek, A., Young, M. A., and Zhang, B. (2024), *The Oxford Handbook of AI Governance* (Oxford University Press).
Burke, E. (1790), *Reflections on the Revolution in France* (James Dodsley).
Caney, S. (2017), 'Political Institutions for the Future: A Fivefold Package', in I. González-Ricoy and A. Gosseries (eds.), *Institutions for Future Generations* (Oxford University Press), 135–155.
Caney, S. (2018), 'Justice and Future Generations', in *Annual Review of Political Science* 21: 475–493.
Caplan, B. (2011), 'The Totalitarian Threat', in N. Bostrom and M. M. Cirkovic (eds.), *Global Catastrophic Risks* (Oxford University Press), 504–518.
Carter, I. (1999), *A Measure of Freedom* (Oxford University Press).
Chan, S. (1997), 'In Search of Democratic Peace: Problems and Promise', in *Mershon International Studies Review* 41: 59–91.
Christiano, T. (2008), *The Constitution of Equality: Democratic Authority and Its Limits* (Oxford University Press).
Cohen, A. and Lee, S. (1986), *Nuclear Weapons and the Future of Humanity: The Fundamental Questions* (Rowman & Littlefield).
Cohen, G. A. (1989), 'On the Currency of Egalitarian Justice', in *Ethics* 99: 906–944.
Cohen, J. (1986), 'An Epistemic Conception of Democracy', in *Ethics* 97: 26–38.
Conly, S. (2016), *One Child: Do We Have a Right to More?* (Oxford University Press).
Cowen, T. (2018), *Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals* (Stripe Press).
Cowen, T. (2023), 'Effective Altruism', University of St Andrews Philosophy, https://www.youtube.com/watch?v=ZzV7ty1DW_c (accessed 12 January 2023).
Cowen, T. and Parfit, D. (1992), 'Against the Social Discount Rate', in P. Laslett and J. S. Fishkin (eds.), *Justice Between Age Groups and Generations* (Yale University Press), 144–168.
Crisp, R. and Pummer, T. (2020), 'Effective Justice', in *Journal of Moral Philosophy* 17: 398–415.
Curran, E. (this volume), 'Longtermism and the Claims of Future People', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Dewey, J. (1927), *The Public and Its Problems* (Henry Holt and Co).
Donaldson, S. and Kymlicka, W. (2011), *Zoopolis: A Political Theory of Animal Rights* (Oxford University Press).
Emanuel, K. (2016), *Climate Science and Climate Risk: A Primer* (MIT).
English, J. (1977), 'Justice between Generations', in *Philosophical Studies* 31: 91–104.
Estlund, D. (2009), *Democratic Authority: A Philosophical Framework* (Princeton University Press).
Fourie, C., Schuppert, F., and Wallimann-Helmer, I. (2015), *Social Equality: On What It Means to Be Equals* (Oxford University Press).
Fried, B. H. (2004), 'Left-Libertarianism: A Review Essay', in *Philosophy & Public Affairs* 32: 66–92.
Friederich, S. and Aebischer, E. (2021), 'At the Precipice Now, in Eternal Safety Thereafter?', in *Metascience* 30: 135–139.
Gardiner, S. M., Caney, S., Jamieson, D., and Shue, H. (2010). *Climate Ethics: Essential Readings* (Oxford University Press).
Garner, R. (2013), *A Theory of Justice for Animals: Animal Rights in a Nonideal World* (Oxford University Press).
Gilabert, P. (2016), 'Justice and Beneficence', in *Critical Review of International Social and Political Philosophy* 19: 508–533.
Giuliano, P. and Nunn, N. (2020), 'Understanding Cultural Persistence and Change', in *The Review of Economic Studies* 88: 1541–1588.
González-Ricoy, I. and Gosseries, A. (2017), *Institutions for Future Generations* (Oxford University Press).
Goodin, R. E. (1980), 'No Moral Nukes', in *Ethics* 90: 417–449.

Goodin, R. E. (1988), 'What Is So Special about Our Fellow Countrymen?', in *Ethics* 98: 663–686.
Goodin, R. E. (1995), *Utilitarianism as a Public Philosophy* (Cambridge University Press).
Goodin, R. E. (2007), 'Enfranchising All Affected Interests, and Its Alternatives', in *Philosophy & Public Affairs* 35: 40–68.
Goodin, R. E. (2016), 'Enfranchising All Subjected, Worldwide', in *International Theory* 8: 365–389.
Goodin, R. E. (2017), 'Duties of Charity, Duties of Justice', in *Political Studies* 65: 268–283.
Goodin, R. E. and Spiekermann, K. (2018), *An Epistemic Theory of Democracy* (Oxford University Press).
Gosseries, A. (2009), 'Three Models of Intergenerational Reciprocity', in A. Gosseries and L. H. Meyer (eds.), *Intergenerational Justice* (Oxford University Press), 119–146.
Gosseries, A. and Meyer, L. H. (2009), *Intergenerational Justice* (Oxford University Press).
Greaves, H. and MacAskill, W. (this volume), 'The Case for Strong Longtermism', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Halstead, J. (2017), 'High Stakes Instrumentalism', in *Ethical Theory and Moral Practice* 20: 295–311.
Hare, C. (2007), 'Voices from Another World: Must We Respect the Interests of People Who Do Not, and Will Never, Exist?', in *Ethics* 117: 498–523.
Heath, J. (2013), 'The Structure of Intergenerational Cooperation', in *Philosophy & Public Affairs* 41: 31–66.
Heath, J. (2017), 'Climate Ethics: Justifying a Positive Social Time Preference', in *Journal of Moral Philosophy* 14: 435–462.
Heikkinen, K. (2022), 'Strong Longtermism and the Challenge from Anti-aggregative Moral Views', GPI Working Paper 5-2022 (Global Priorities Institute, Oxford University).
Herre, B. and Roser, M. (2013), 'Democracy', *Our World in Data*, https://ourworldindata.org/democracy (accessed March 2023).
Heyd, D. (2009), 'A Value or an Obligation? Rawls on Justice to Future Generations' in A. Gosseries and L. H. Meyer (Eds.), *Intergenerational Justice* (Oxford University Press), 167–188.
Hirose, I. (2022), *The Ethics of Pandemics: An Introduction* (Routledge).
Hong, L. and Page, S. E. (2004), 'Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers', in *Proceedings of the National Academy of Sciences* 101: 16385–16389.
Hourdequin, M. and Wong, D. B. (2021), 'Confucianism and Intergenerational Ethics', in S. M. Gardiner (ed.), *The Oxford Handbook of Intergenerational Ethics* (Oxford University Press).
Inglehart, R., Foa, R., Peterson, C., and Welzel, C. (2008), 'Development, Freedom, and Rising Happiness: A Global Perspective (1981–2007)', in *Perspectives on Psychological Science* 3: 264–285.
Jackson, F. (1991), 'Decision-Theoretic Consequentialism and the Nearest and Dearest Objection', in *Ethics* 101: 461–482.
John, T. (2022), *Chronopolitanism: Political Institutions and the Distant Future*, PhD thesis, Rutgers University.
John, T. and MacAskill, W. (2021), 'Longtermist Institutional Reform', in N. Cargill and T. John (eds.), *The Long View: Essays on Policy, Philanthropy, and the Long-term Future* (FIRST), 44–60.
Jonas, H. (1979), *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation* (Suhrkamp).
Jonas, H. (1985), *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (University of Chicago Press).
Jones, C. I. (2022), 'The End of Economic Growth? Unintended Consequences of a Declining Population', in *American Economic Review* 112: 3489–3527.
Kagan, S. (2011), 'Do I Make a Difference?', in *Philosophy & Public Affairs* 39: 105–141.
Katz, C. (2017), 'Neorepublicanism and the Domination of Posterity', in *Ethics, Policy & Environment* 20: 294–313.
Kavka, G. S. (1987), *Moral Paradoxes of Nuclear Deterrence* (Cambridge University Press).
Knight, J. and Johnson, J. (2011), *The Priority of Democracy: Political Consequences of Pragmatism* (Princeton University Press).
Kosonen, C. (2023), 'Tiny Probabilities and the Value of the Far Future', GPI Working Paper Series 1-2023 (Global Priorities Institute, Oxford University).
Kramer, M. H. (2003), *The Quality of Freedom* (Oxford University Press).
Kristjánsson, K. (1996), *Social Freedom: The Responsibility View* (Cambridge University Press).
Krznaric, R. (2021), *The Good Ancestor: How to Think Long Term in a Short-Term World* (W. H. Allen).
Kumar, R. (2015), 'Risking and Wronging', in *Philosophy and Public Affairs* 43: 27–51.
Kumar, R. (2018), 'Risking Future Generations', in *Ethical Theory and Moral Practice* 21: 245–257.
Landemore, H. (2012), *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many* (Princeton University Press).
Lechterman, T. M. (2021), *The Tyranny of Generosity: Why Philanthropy Corrupts Our Politics and How We Can Fix It* (Oxford University Press).

Lewis, D. (1986), 'Buy Like a MADman, Use Like a NUT', in *Qq* 6: 5–8.
Lippert-Rasmussen, K. (2015), Luck Egalitarianism (Bloomsbury Publishing).
Lippert-Rasmussen, K. (2018), *Relational Egalitarianism: Living as Equals* (Cambridge University Press).
List, C. and Goodin, R. E. (2001), 'Epistemic Democracy: Generalizing the Condorcet Jury Theorem', in *Journal of Political Philosophy* 9: 277–306.
Locke, J. (1690/1990), *Second Treatise on Civil Government*, ed. C. B. MacPherson (Hackett).
Lovett, F. (2010), *A General Theory of Domination and Justice* (Oxford University Press).
MacAskill, W. (2022), *What We Owe the Future* (Hachette).
Mahoney, J. (2000), 'Path Dependence in Historical Sociology', in *Theory and Society* 29: 507–548.
Marglin, S. A. (1963), 'The Social Rate of Discount and the Optimal Rate of Investment', *The Quarterly Review of Economics* 77: 95–111.
Matheny, G. (2002), 'Expected Utility, Contributory Causation, and Vegetarianism', in *Journal of Applied Philosophy* 19: 293–297.
Mbonda, E.-M. and Ngosso, T. (2021), 'Intergenerational Justice: An African Perspective', in S. M. Gardiner (ed.), *The Oxford Handbook of Intergenerational Ethics* (Oxford University Press).
Meyer, L. (2021), 'Intergenerational Justice', in E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, Summer 2021, https://plato.stanford.edu/entries/justice-intergenerational/ (accessed March 2023).
Meyer, L. and Roser, D. (2009), 'Enough for the Future', in A. Gosseries and L. H. Meyer (eds.), *Intergenerational Justice* (Oxford University Press), 219–248.
Mill, J. S. (1866), 'Army—Regiments In India', in *House of Commons Hansard*, vol. 182 (UK Parliament). https://hansard.parliament.uk//Commons/1866-04-17/debates/d4ba0459-2d9f-468f-b589-7321ecc1dfb3/Army—RegimentsInIndia(accessed March 2023).
Miller, D. (1983), 'Constraints on Freedom', in *Ethics* 94: 66–86.
Miller, D. (1997), 'Equality and Justice', in *Ratio* 10: 222–237.
Miller, D. (2020), 'Reconceiving the Democratic Boundary Problem', in *Philosophy Compass* 15: 1–9.
Miller, D. (2021), 'Justice', in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2021, https://plato.stanford.edu/archives/fall2021/entries/justice/
Mittiga, R. (2022), 'Political Legitimacy, Authoritarianism, and Climate Change', in *American Political Science Review* 116: 998–1011.
Moellendorf, D. (2015), 'Climate Change Justice', in *Philosophy Compass* 10: 173–186.
Mogensen, A. L. (2021), 'Moral Demands and the Far Future', in *Philosophy and Phenomenological Research* 103: 567–585.
Mogensen, A. L. (2022), 'The Only Ethical Argument for Positive ? Partiality and Pure Time Preference', in *Philosophical Studies* 179: 2731–2750.
Mogensen, A. L. (this volume), 'Would a World without Us Be Worse? Clues from Population Axiology', in H. Greaves, J. Barrett, and D. Thorstad (eds.), *Essays on Longtermism* (Oxford University Press).
Narveson, J. (1973), 'Moral Problems of Population,' in *The Monist* 57: 62–86.
Norcross, A. (2004), 'Puppies, Pigs, and People: Eating Meat and Marginal Cases', in *Philosophical Perspectives* 18: 229–245
Nozick, R. (1974), *Anarchy, State, and Utopia* (Basic Books).
Office of Management and Budget. (2021), *Budget of the U.S. government—Fiscal Year 2022* (U.S. Government Publishing Office).
O'Neill, M. (2008), 'What Should Egalitarians Believe?', in *Philosophy & Public Affairs* 36: 119–156.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Hachette).
Ortiz-Ospina, E. (2019), 'Does Democracy Lead to Better Health?', *Our World in Data*, https://ourworldindata.org/democracy-health (accessed March 2023).
Orviska, M., Caplanova, A., and Hudson, J. (2014), 'The Impact of Democracy on Well-being', in *Social Indicators Research* 115: 493–508.
Otsuka, M. (2003), *Libertarianism without Inequality* (Oxford University Press).
Owen, A. L., Videras, J., and Willemsen, C. (2008), 'Democracy, Participation, and Life Satisfaction', in *Social Science Quarterly* 89: 987–1005.
Page, S. E. (2008), *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies - New Edition* (Princeton University Press).
Parfit, D. (1991), *'Equality or Priority?', The Lindley Lecture* (University of Kansas).
Parijs, P. V. (1997), *Real Freedom for All: What (if Anything) Can Justify Capitalism?* (Clarendon Press).
Pettit, P. (2012), 'The Inescapability of Consequentialism', in U. Heuer and G. Lang (eds.), *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams* (Oxford University Press), 41–70.
Pettit, P. (2014), *Just Freedom: A Moral Compass for a Complex World* (W. W. Norton & Company).

Pierson, P. (2000), 'Increasing Returns, Path Dependence, and the Study of Politics', in *American Political Science Review* 94: 251–267.
Porter, J. and Gibbons, A. (2024), 'Existential Risk and Equal Political Liberty', in *Asian Journal of Philosophy* 3: 1–26.
Railton, P. (1984), 'Alienation, Consequentialism, and the Demands of Morality', in *Philosophy & Public Affairs* 13: 134–171.
Rawls, J. (1971), *A Theory of Justice* (Harvard University Press).
Rawls, J. (2001), *The Law of Peoples* (Harvard University Press).
Rawls, J. (2005), *Political Liberalism* (Columbia University Press).
Reich, R. (2018), *Just Giving: Why Philanthropy Is Failing Democracy and How It Can Do Better* (Princeton University Press).
Rendall, M. (2021), 'Nuclear Deterrence—Another Perfect Storm', in S. M. Gardiner (ed.), *The Oxford Handbook of Intergenerational Ethics* (Oxford University Press).
Riedener, S. (2022), 'Human Extinction from a Thomist Perspective', in D. Roser, S. Riedener, and M. Huppenbauer (eds.), *Effective Altruism and Religion: Synergies, Tensions, Dialogue* (Nomos), 187–210.
Rubin, O. (2009), 'The Merits of Democracy in Famine Protection – Fact or Fallacy?' in *The European Journal of Development Research* 21: 699–717.
Russell, J. (2024), 'On Two Arguments for Fanaticism', in Nous 58: 565–595.
Russell, J. (2022), 'Problems for Impartiality', *The Parfit Memorial Lecture* (16 June 2022), https://globalprioritiesinstitute.org/parfit-memorial-lecture-jeffrey-sanford-russell-problems-for-impartiality/ (accessed March 2023).
Sangiovanni, A. (2007), 'Global Justice, Reciprocity, and the State', in *Philosophy & Public Affairs* 35: 3–39.
Saunders, B. (2012), 'Defining the Demos', in *Politics, Philosophy and Economics* 11: 280–301.
Saunders-Hastings, E. (2022), *Private Virtues, Public Vices: Philanthropy and Democratic Equality* (University of Chicago Press).
Scanlon, T. M. (1998), *What We Owe to Each Other* (Harvard University Press).
Scanlon, T. M. (2018), *Why Does Inequality Matter?* (Oxford University Press).
Scheffler, S. (2018), *Why Worry About Future Generations?* (Oxford University Press).
Schemmel, C. (2021), *Justice and Egalitarian Relations* (Oxford University Press).
Schmidt, A. T. (2016), 'Abilities and the Sources of Unfreedom', in *Ethics* 127: 179–207.
Schmidt, A. T. (2018a), 'Domination without Inequality? Mutual Domination, Republicanism, and Gun Control', in *Philosophy & Public Affairs* 46: 175–206.
Schmidt, A. T. (2018b), 'Persons or Property – Freedom and the Legal Status of Animals', in *Journal of Moral Philosophy* 15: 20–45.
Schmidt, A. T. (2025), 'The Freedom of Future People', in *Philosophy & Public Affairs* 53 (2): 197–214. https://doi.org/10.1111/papa.12283.
Schmidt, A. T. (2024a), 'Consequentialism, Collective Action, and Blame,' in *Journal of Moral Philosophy* 22 (1–2): 183–215. https://doi.org/10.1163/17455243-20244215.
Schmidt, A. T. (2024b), 'Egalitarianism across Generations', in *Utilitas* 36: 242–264.
Schmidt, A. T. (2023a), 'Consequentialism and the Role of Practices in Political Philosophy', in *Res Publica* 30: 429–450.
Schmidt, A. T. (2022c), 'From Relational Equality to Personal Responsibility', in *Philosophical Studies* 179: 1373–1399.
Schmidt, A. T. (2022d), 'Freedom in Political Philosophy', in *Oxford Research Encyclopedia of Politics* (Oxford University Press), https://doi.org/10.1093/acrefore/9780190228637.013.2022
Schmidt, A. T. and Bengtson, A. (2021), 'Future People and the Boundaries of Republican Democracy' (unpublished manuscript).
Schmidt, A. T. and Juijn, D. (2024), 'Economic Inequality and the Long-Term Future', in *Politics, Philosophy & Economics* 23 (1): 67–99. https://doi.org/10.1177/1470594X231178502.
Schmidtz, D. (2011), 'Nonideal Theory: What It Is and What It Needs to Be', in *Ethics* 121: 772–796.
Segall, S. (2016), 'Incas and Aliens: The Truth in Telic Egalitarianism', in *Economics and Philosophy* 32: 1–19.
Sen, A. (1982), *Poverty and Famines: An Essay on Entitlement and Deprivation* (Oxford University Press).
Sen, A. (1999), *Development as freedom* (Knopf).
Sevilla, J. (2022), 'Persistence: A Critical Review', in *What We Owe the Future - Supplementary Material*,https://whatweowethefuture.com/wp-content/uploads/2023/06/Persistence-a-critical-review.pdf (accessed March 2023).
Shahar, D. C. (2015), 'Rejecting Eco-authoritarianism, Again', in *Environmental Values* 24: 345–366.
Shklar, J. N. (1990), *The Faces of Injustice* (Yale University Press).
Shulman, C. and Thornley, E. (this volume), 'How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role', in H. Greaves, J. Barrett, and D. Thorstad (eds.) *Essays on Longtermism* (Oxford University Press).

Simmons, A. J. (2001), *Justification and Legitimacy: Essays on Rights and Obligations* (Cambridge University Press).
Simmons, A. J. (2010), 'Ideal and Nonideal Theory', in *Philosophy & Public Affairs* 38: 5–36.
Singer, P. (1972), 'Famine, Affluence, and Morality', in *Philosophy & Public Affairs* 1: 229–243.
Singer, P. (1980), 'Utilitarianism and Vegetarianism', in *Philosophy & Public Affairs* 9: 325–337.
Skinner, Q. (2012), *Liberty before Liberalism* (Cambridge University Press).
Smith, P. T. (2013), 'The Intergenerational Storm: Dilemma or Domination', in *Philosophy and Public Issues* 3: 207–244.
Sobel, D. (2020), 'Understanding the Demandingness Objection', in D. W. Portmore (ed.), *The Oxford Handbook of Consequentialism* (Oxford University Press), 221–238.
Song, S. (2012), 'The Boundary Problem in Democratic theory: Why the Demos Should be Bounded by the State', in *International Theory* 4: 39–68.
Steiner, H. (1994), *An Essay on Rights* (Wiley).
Stemplowska, Z. (2013), 'Rescuing Luck Egalitarianism', in *Journal of Social Philosophy* 44: 402–419.
Stern, N. (2006), *The Economics of Climate Change: The Stern Review* (Cambridge University Press).
Tännsjö, T. (2007), 'Future People, the All Affected Principle, and the Limits of the Aggregation Model of Democracy', in T. Rønnow-Rasmussen, B. Petersson, J. Josefsson, and D. Egonsson (eds.), *Hommage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, https://www.fil.lu.se/hommageawlodek/index.htm (accessed March 2023).
Tarsney, C. J. (2023), 'The Epistemic Challenge to Longtermism', in *Synthese* 201: 195.
Temkin, L. S. (1995), 'Justice and Equality: Some Questions About Scope', in *Social Philosophy and Policy* 12: 72–104.
Thomas, T. (2023), 'The Asymmetry, Uncertainty and the Long Term', in *Philosophy and Phenomenological Research* 107/2: 470–500.
Thorstad, D. (2023), 'High Risk, Low Reward: A Challenge to the Astronomical Value of Existential Risk Mitigation', in Philosophy & Public Affairs 51: 373–412.
Martinez, E. and Winter, C. (this volume), 'The Intuitive Appeal of Legal Protection for Future Generations', in H. Greaves, D. Thorstad, and J. Barrett (eds.), *Essays on Longtermism* (Oxford University Press).
United Nations. (2021), *Our Common Agenda – Report of the Secretary-General* (United Nations).
Valentini, L. (2011a), 'Coercion and (Global) Justice', in *American Political Science Review* 105: 205–220.
Valentini, L. (2011b), *Justice in a Globalized World: A Normative Framework* (Oxford University Press).
Vallentyne, P. and Steiner, H. (2000), *Left-Libertarianism and Its Critics: The Contemporary Debate* (Palgrave Macmillan).
Veenhoven, R. (2000), 'Freedom and Happiness: A Comparative Study in Forty-Four Nations in the Early 1990s', in E. Diener and E. M. Suh (eds.), *Culture and Subjective Well-being* (MIT Press), 257–288.
Vercelli, A. (1998), 'Sustainable Development and the Freedom of Future Generations', in G. Chichilnisky, G. Heal, and A. Vercelli (eds.), *Sustainability: Dynamics and Uncertainty* (Springer Netherlands), 171–187.
Vidiella, G. and García Valverde, F. (2021), 'Buen Vivir: A Latin American Contribution to Intra- and Intergenerational Justice', in S. M. Gardiner (ed.), *The Oxford Handbook of Intergenerational Ethics* (Oxford University Press).
Weitzman, M. L. (1998), 'Why the Far-Distant Future Should Be Discounted at Its Lowest Possible Rate', *Journal of Environmental Economics and Management* 36: 201–208.
Wilkinson, H. (2022), 'In Defense of Fanaticism', in *Ethics* 132: 445–477.
Woodard, C. (2019), *Taking Utilitarianism Seriously* (Oxford University Press).
Young, I. M. (1990), *Justice and the Politics of Difference* (Princeton University Press).
Ziliotti, E. (2020), 'Democracy's Value: A Conceptual Map', in *Journal of Value Inquiry* 54: 407–427.

27

Retrospective Accountability

A Mechanism for Representing Future Generations

Tyler M. John

A number of recent writers have argued that the obligations of modern states to future generations may far outstrip their obligations to their present citizens, given the vast number of people who will exist in the future and whose livelihoods depend on our actions (Beckstead 2013; Greaves and MacAskill 2021; Tarsney 2019). And yet modern states do precious little on behalf of future generations, opting to allow and incentivise destructive practices such as the widespread burning of fossil fuels, while failing to take preventative measures that could deter global pandemics and other catastrophes.

The state is plagued with problems of political short-termism: the excessive priority given to near-term benefits at the cost of future ones (González-Ricoy and Gosseries 2016). By the accounts of many political scientists and economists, political leaders rarely look beyond the next 2–5 years and into the problems of the next decade. There are many reasons for this, from time preference (Frederick et al. 2002; Jacobs and Matthews 2012) to cognitive bias (Weber 2006; Johnson and Levin 2009; Caney 2016) to re-election incentives (Mayhew 1974; Tufte 1978; Arnold 1990; Binder 2006),[1] but all involve foregoing costly action in the short term (e.g. increasing taxes, cutting benefits, imposing regulatory burdens) that would have larger moderate- to long-run benefits. Such behaviour is detrimental not only to the generations of people who are to come, but also to the large number of existing citizens who still have much of their lives left to lead.

A small literature in philosophy, economics, and political science considers what modern governments might do to ameliorate short-termism and incentivise states to adequately prioritise the interests of future generations in policymaking (see especially Caney 2016; González-Ricoy and Gosseries 2016; John and MacAskill 2021). One core unsolved problem in this literature is the problem of developing incentive mechanisms for present policymakers to promote the welfare of future generations. For most disenfranchised groups, simply giving them the vote and ensuring a functional democratic political system ensures that their preferences are represented in government. But future generations cannot cast a ballot—so how can we ensure that governments adequately protect them?

I have previously criticised extant attempts to solve this problem (John 2020; 2023). This chapter is my attempt to develop a successful strategy for incentivising the present generation to act on behalf of future generations: building a novel *futures assembly* which is explicitly incentivised to promote the interests of future generations. The assembly I propose is governed by citizens randomly selected from among the populace, who are rewarded in the

[1] For a contrary view, see Beck (1982).

Tyler M. John, *Retrospective Accountability* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Tyler M. John 2025. DOI: 10.1093/9780191979972.003.0027

future to the extent that they are successful in promoting the welfare of future generations. In section 1, I outline a novel futures assembly and detail the two essential features of its mechanism design: sortition (random selection of assembly members) and retrospective liability (rewarding assembly members in the future, based on their past performance vis-à-vis safeguarding the future). Section 2 elaborates the function of the assembly, including the duties and powers that it should have. The assembly's structure is spelled out in section 3, including the constitution of the assembly, the role of experts, and its research and media departments. Section 4 briefly discusses how the assembly should interact with the rest of government. I conclude with some reflections on the epistemic status of the institution I have proposed. In particular, I consider the extent to which this mechanism can extend today's time horizons—whether it can plausibly improve the world in not merely 1,000 years, but also 1 million years hence.

1 A forward-looking assembly

Two key questions for any proposed political institution are: (i) Who serves? (ii) What are their incentives? The body I propose is populated by random selection of political decision-makers from among the populace, who are educated on issues affecting future generations prior to making political decisions, to ensure that the people who serve are unelected, informed, non-polarised, and representative. It is incentivised by retrospective accountability, a mechanism on which decision-makers are rewarded years after their tenure based on the effects of their decisions over that period, to ensure that the people who serve are incentivised to promote the well-being of future generations. In this section I articulate and defend these two core features of the proposed representative mechanism.

1.1. Sortition

'Citizens assemblies' have been employed for consultation and information-gathering purposes throughout the world. These randomly selected groups of citizens provide deliberative and non-binding advice to the government in consultation with recognised experts. One of the most high-profile initiatives was Ireland's 100-member Citizens' Assembly, which was established in 2016 and tasked with considering questions related to abortion, fixed-term parliaments, referenda, population aging, climate change, and gender equality. The deliberations of the Irish assembly provoked a referendum to remove Ireland's constitutional ban on abortion and substantially shaped Ireland's Climate Action Plan (Coleman et al. 2019).

The success of the Irish assembly and of citizens' assemblies around the world reveals the promise of citizens' assemblies tasked with the explicit mandate to represent future generations, or 'futures assemblies'. A general futures assembly, constituted by a stratified random sample of the general population, would have numerous features that predict success at combating short-termism (John and MacAskill 2021). Being an unelected and publicly funded body, a futures assembly would be insulated from the election and fundraising incentives that pressure elected officials to focus on near-term,

visible issues. Being randomly selected, it would be statistically representative of the general population. And citizens' assemblies have a demonstrated aptitude in 'laboratory' and real-world experiments for reducing the deleterious effects of partisanship on careful, long-term deliberation (Fishkin and Luskin 2005; List et al. 2013; Fishkin et al. 2017). In the most recent major assembly, the Climate Assembly UK, 98% of assembly members claimed to have understood almost everything that those in their deliberation groups had said, and 94% felt respected by their fellow participants under disagreement (with none feeling disrespected) (Climate Assembly UK 2020). Finally, citizens' assemblies are more informed than ordinary voters due to their deliberations with experts, reducing the deleterious effect of policy uncertainty on short-term policy support (cf. Jacobs and Matthews 2012).

A general futures assembly may need no incentive to reflect carefully on the interests of future generations beyond an explicit mandate to do so. Some limited evidence from the Kochi University Research Institute for Future Design suggests that when parents are explicitly asked to cast votes on behalf of their children, they vote for different parties than they normally would vote for in a sizable minority of cases (Aoki and Vaithianathan 2012). This is a promising sign that those who are asked explicitly to represent other generations do not simply use the opportunity to promote their own agenda, but rather aim to promote the interests of the relevant generation, and thereby adopt longer time horizons for political decision-making. This is further supported by evidence that actors within institutions tend to be compelled to follow norms that are consistent with the established culture of their institution (Goodin 1995; Steiner et al. 2005: 127; MacKenzie 2016).

1.2 Retrospective accountability

An explicit mandate may be sufficient to motivate futures assembly members to give substantial consideration to future generations in their recommendations. But there are a number of reasons why their consideration of future generations may be inadequate, including public pressure, value misalignment, value drift, weakness of will, intrinsic time discounting (or impatience), and corruption. Moreover, empirical evidence that political actors follow the established rules of their institutions is limited, and we do not know the universality of its scope or the strength of its implications. We can have much greater confidence in the decisions of a randomly selected futures assembly if it is directly incentivised to promote the welfare of even distant future generations.

An underexplored mechanism for aligning incentives with the interests of future generations involves *retrospective accountability*. The central problem of representing the interests of future generations in government is that of making political actors accountable to future generations. Future generations cannot vote in our elections, nor can they sanction or protest the decisions of their forebears. Retrospective accountability solves the accountability problem by rewarding policy-makers years into the future in proportion to the effects of their policy on the long run. A simple mechanism of retrospective accountability would involve empowering a body of future auditors—say, 30 years from now—to decide on the pension bonus of the decision-makers today based on how successfully these decision-makers promote the interests of future people. This would provide decision-makers today with a positive financial incentive to look to the future—at least 30 years from now—when

making any decisions. Such a mechanism would yield a significant advance on the time horizons of present institutions.[2]

There are a variety of retrospective accountability mechanisms that can be employed by governments. First, governments can place an age limit on the relevant class of political decision-makers and reward people late in life. If the age limit is 30–40 years, and they are rewarded around the age of 60, this could extend political time horizons by 20–30 years. Second, governments can choose decision-makers who are parents or otherwise connected to children, and reward these children later in life based on the decisions of the parents. This could extend political time horizons by 40 years, or even longer. Third, governments can find other rewards that incentivise policymakers even when the rewards are given out after they have died, such as by promising to further their projects, build their communities, or improve their legacy. This could in principle extend the time horizons of policymakers indefinitely (discounted by the likelihood that these commitments will be fulfilled), but we do not at this time know how strongly such promises could motivate policymakers, or which such mechanisms are most powerful. Fourth, governments can employ retrospective accountability mechanisms *iteratively*, selecting a sequence of policymakers who each decide on the later bonus of a previous generation, with each of their own future bonuses tied to the opinions of the following generation. I think that such an iterated accountability mechanism is promising, and it warrants detailed analysis.

1.3 Iterated retrospective accountability

One possible approach would exploit strategic iteration of this mechanism to extend the time horizons of government far into the future. On the iterated variant, the future auditors who decide on the future bonuses of present decision-makers *themselves* face a financial incentive to look again into the future. For their own financial situation will be tied to the evaluations of the *next* generation of auditors, who will determine their pension bonuses. To get a nice retirement bonus, future auditors have an incentive to evaluate present decision-makers in accordance with the preferences of the next generation of auditors, and so present decision-makers have an incentive to satisfy the preferences of the auditors two generations—60 years—from now. And so iterated, until we have extended the horizons of government to the longest time period relevant for political decision-making. On the simplest implementation of such accountability measures, the assemblies are also the auditing bodies: each assembly decides on the bonus of the assembly 30 years prior.

We should give special attention to two subspecies of iterated accountability mechanisms, differentiated by whether there is a known period when the iterated mechanism terminates. In the 'Final Tribunal of Justice' model, there is a known final auditor in the series whose judgments are not incentivised by a subsequent auditor's retrospective rewards. On the 'Infinite Justice' model, there is no known final auditor; the series continues indefinitely,

[2] Retrospective accountability has a similar incentives structure to electoral democracy. As some contemporary theorists of democratic accountability claim, democracies make policymakers accountable to the public through a process of 'retrospective voting' (Key 1966; Fiorina 1977; Ferejohn 1986; Rogoff 1990; Banks and Sundaram 1993; Fearon 1999). However, unlike electoral democracy, retrospective democracy has much longer time horizons, and offers a much tighter connection between decisions and rewards (assembly members are audited *after* they leave office).

with some positive probability of termination at every period, due to either repeal or social collapse. In this section I show that both subspecies of iterated accountability mechanism can succeed at incentivising political decision-makers to significantly extend their time horizons, under certain conditions.

1.3.1 The Final Tribunal of Justice

In the Final Tribunal of Justice model, the iterated accountability mechanism has a known final auditor. This auditor is not evaluated by any further auditors, and so has no incentive from the mechanism to cooperate with the rules of the scheme: namely, to reward the previous generation to the extent that they chose the 'optimal strategy', that is, chose the optimal policies and adopted a reward strategy that optimally incentivised the previous generation to choose optimal policies and rewards.

For this reason, the Final Tribunal of Justice succeeds in incentivising the chain of auditors to choose the optimal strategy if and only if each generation n is sufficiently confident that generation $n + 1$ is sufficiently confident that generation $n + 2$ is sufficiently confident that . . . generation $n + i$, the final generation, will be motivated to choose the optimal strategy.[3] The precise degree of confidence required depends on the auditors' own preferences about whether to cooperate or defect, and the size of the reward for cooperation.[4] For ease of exposition, I'll assume that the degree of confidence required by each generation is full belief.

Suppose that a generation n believes that generation $n + 1$ believes that generation $n + 2$ believes that . . . generation $n + i$, the final generation, will be motivated to choose the optimal strategy. Then generation n can use backwards induction to see that they, too, ought to use the optimal strategy. Using the optimal strategy involves rewarding the previous generation *iff* they chose the optimal strategy. So, if the final auditor n chooses the optimal strategy, they will reward the previous generation $n - 1$ *iff* $n - 1$ chooses the optimal strategy. If $n - 1$ believes this, they will be motivated to choose the optimal strategy. And if $n - 2$ believes that $n - 1$ will be motivated to choose the optimal strategy, they also believe that $n - 1$ will be motivated to reward the previous generation $n - 2$ *iff* $n - 2$ chooses the optimal strategy. So $n - 2$ ought to infer, via backwards induction, that they will be rewarded if they choose the optimal strategy. And so iterated, until we reach the first auditor, who too can infer, via backwards induction, that they will be rewarded if they choose the optimal strategy.

By contrast, suppose that some generation n does *not* believe that generation $n + 1$ believes that generation $n + 2$ believes that . . . generation $n + i$, the final generation, will be motivated to choose the optimal strategy. Then n cannot infer that $n + 1$ will reward n for choosing the optimal strategy. In fact, if n is sufficiently confident that it is *not* the case that generation $n + 1$ believes that generation $n + 2$ believes that . . . generation $n + i$, the final generation, will be motivated to choose the optimal strategy, then n can employ backwards

[3] Note that this condition is substantially weaker than common knowledge.

[4] The precise level of confidence required can be derived from the auditor's own preferences about whether to cooperate or defect and the size of the reward for cooperation. If the auditor is intrinsically motivated to cooperate, the auditor does not need to be very confident that the subsequent auditor will cooperate. If the auditor is intrinsically motivated to defect, they need to be more confident that the subsequent auditor will cooperate, or the size of the reward for cooperating needs to be larger. In general, as long as each auditor believes that there is a better-than-chance likelihood of being rewarded for cooperation, cooperation can be sustained for an arbitrarily large reward.

induction to see that they will *not* be rewarded for choosing the optimal strategy. Consider a random generation $n + k$ such that n is confident that generation $n + 1$ believes that generation $n + 2$ believes that . . . generation $n + k$ does *not* believe that the subsequent generation will be motivated to choose the optimal strategy. Then n can infer that $n + 1$ can infer that $n + 2$ can infer that . . . $n + k - 1$ can infer that $n + k$ will have no incentive to reward $n + k - 1$ for choosing the optimal strategy. And so n can infer that $n + 1$ can infer that $n + 2$ can infer that . . . $n + k - 1$ will have no incentive to choose the optimal strategy. And so n can infer that $n + 1$ can infer that $n + 2$ can infer that . . . $n + k - 1$ will have no incentive to reward $n + k - 2$ for choosing the optimal strategy. And so on, finally licensing the inference that $n + 1$ has no incentive to choose the optimal strategy, providing no incentive, in turn, to n to choose the optimal strategy.

So, the Final Tribunal of Justice succeeds in incentivising the chain of auditors to choose the optimal strategy if and only if each generation n is sufficiently confident that generation $n + 1$ is sufficiently confident that generation $n + 2$ is sufficiently confident that . . . generation $n + i$, the final generation, will be motivated to choose the optimal strategy. This condition would be trivially met if there were common knowledge that the final generation will be motivated to choose the optimal strategy. So one way to ensure cooperation is to make it public that the final generation *is* motivated to choose the optimal strategy. This could happen if the scheme were committed to the choice of a final generation with the right motivational profile—perhaps people who are deeply motivated to cooperate with the scheme, or people who are motivated by justice for their own generation—and making it public that this is the case. In some light this does not seem particularly difficult. By default, it would be surprising if the final generation were *not* significantly motivated to judge the previous generation harshly if the previous generation had failed them, and to judge them favourably if the previous generation had helped them. Juries have no external motivation to judge court cases aptly, but their sense of justice and the rules of the scheme ensure that juries work reasonably well. However, if it were unclear whether the final auditor would be properly motivated, the scheme would fall apart.

1.3.2 Infinite Justice

On the Infinite Justice model, the iterated accountability mechanism has no known final auditor. So each auditor in the series has incentive from the mechanism to cooperate with the rules of the scheme and to reward the previous generation to the extent that they chose the optimal strategy. In the Final Tribunal of Justice model, each generation must believe that the last generation will be intrinsically motivated to follow the scheme. But in the Infinite Justice model no such guarantee is required.

The Infinite Justice model assumes that policymakers are motivated to choose short-termist policy. In the absence of intervention, then, there is one unique perfect subgame equilibrium: policymakers will always choose the short-termist policy and cannot rationally deviate from the choice of a short-termist policy. By offering a bonus to policymakers that depends on their policy choice, the Infinite Justice model allows for multiple perfect subgame equilibria depending on their expectations about what future auditors will reward. If a generation of auditors expects future auditors to reward policy with time horizon H, and they are rational, then this generation of auditors will in fact choose policy with time horizon H, under the specific conditions outlined in subsection 1.3.3.

The Infinite Justice model can sustain cooperation on ideal policy because, so long as the per-generation probability that the mechanism will collapse is sufficiently low, the auditors are unable to determine which generation is the last generation. No auditor can be confident that they or the next generation will be the last generation, and so they cannot be confident that they or the next generation have no incentive to cooperate with the scheme. As long as there is a sufficiently high chance that there will be two subsequent generations, an auditor can expect to reap rewards for choosing the optimal policy.

It is provable that the Infinite Justice model can sustain cooperation on a set of policy goals across many generations.[5] But it is compatible with multiple perfect subgame equilibria. A corrupt line of auditors could collectively coordinate to give the previous generation a bonus no matter what. As long as the auditors expect the next few generations to play the same game, rewarding the previous auditors no matter what, they could continue to adopt the short-termist policies they prefer. So we need to have reason to expect that they will coordinate on *long*termist policy goals, such that most generations of auditors expect future auditors to reward them for adopting a long time horizon. In particular, we need to have reason to expect that longtermist policy is a salient Schelling Point on which each generation can coordinate.

In the real world, we should expect an equilibrium to emerge around long rather than short time horizons. First, as discussed previously, political actors find following the mandates of their office intrinsically rewarding, and this office will mandate adopting long time horizons. This provides direct incentive for policymakers to choose a long time horizon, but it also gives each generation reason to expect that future generations of policymakers will also choose a long time horizon, and so it gives them reason to expect that they will be rewarded for doing the same. Second, only a cooperative equilibrium with long time horizons will be communicated to policymakers, and so by default we should expect that they will settle on this equilibrium. And third, in any sensible state, a corrupt and persistent scheme of defection from the purposes of the assembly will be repealed. And given that the bonus of each generation depends on the system *not* being repealed before the audit, a rational futures assembly will choose not to defect, given that doing so would eliminate their reward. That is, though the rational choice model assumes that the probability that the futures assembly will be repealed is independent of choices made by the assembly, this is clearly not true. If the first assembly makes a series of short-termist policy decisions and the second assembly rewards them for it, the political contemporaries of the second assembly will repeal the assembly for failing to abide by the rules of the office, and for failing to promote their interests. So a rational such assembly will not make a series of short-termist policy decisions.

The arational pressures towards intergenerational cooperation due to intrinsic motivation to follow the norm and the narrative set by the office, as well as external pressures of repeal against a corrupt office, should imply a cooperative equilibrium around long time horizons, even where participants in the scheme deviate from game-theoretic rationality. Without small-scale intergenerational cooperation experiments it is difficult to know how successful a scheme of Infinite Justice would be in practice. But the underpinning game theory illuminates the model as a promising mechanism for further investigation and

[5] John (2022).

highlights the institutional features that are relevant to such a system working in the real world, as I next discuss.

1.3.3 Conditions of success

Mechanisms using either the Final Tribunal of Justice model or the Infinite Justice model need to satisfy some conditions if they are to succeed. Specifically, the expected value of the bonus to the auditors—discounting for the chance that they die or the next generation defects or the next generation knows when the mechanism will be repealed—must be higher than the expected intrinsic reward of the next generation's giving them the bonus. If, in either model, this condition fails to hold, then the mechanism cannot sustain cooperation.

For the condition to hold, a few things must happen. First, and roughly, auditors cannot be nearly as excited to give money to the previous generation as they are to receive money themselves. That is, they need to value their own money more than they value the previous generation's money, at least by several times. Fortunately, it is generally true that people value their own money at least several times more than they value strangers' money. Second, auditors cannot have too high an expectation of dying before they receive and can benefit from the money (or they must value passing it on to their kin about as much as they value receiving it). So the temporal distance between auditing and audited generations cannot be too far. Third, auditors cannot be too confident that the office will be repealed before the next auditing cycle, or that the next auditor will know when the office will be repealed. For this reason, institutional longevity should be a key desideratum of retrospective accountability mechanisms. And fourth, auditors need to have sufficiently high confidence that they can do their job well enough to receive a bonus. This means the conditions for the bonus cannot be overly demanding, and that futures assembly members need tools to be confident that they can do their jobs well.

This final criterion of success entails some further knowledge conditions. In particular, a majority of auditors must be sufficiently confident that the next generation will cooperate, and so they must be sufficiently confident that the next generation understands the optimal strategy profile and the reasons for cooperation. If they have too high an expectation of ignorance, they will place a low expected value on cooperation, since the link between their cooperation and receiving their bonus will be tenuous.

In this section I've proposed a bare-bones futures assembly that employs two core design features: sortition and retrospective accountability. Sortition functions to select a statistically representative unelected body of citizens to serve in the assembly. Retrospective accountability functions to solve a kind of intergenerational principal-agent problem, making the incentives of present decision-makers tied to the choices of future generations. The discussion has raised two key desiderata for a successful futures assembly employing these design features:

1. **The Epistemological Criterion**: futures assemblies must have tools for solving difficult epistemological problems in order to have the confidence that they can do their jobs well enough to be rewarded, and so that later assemblies will have the capacity to correctly evaluate them.
2. **The Longevity Criterion**: futures assemblies need to be highly resilient, with sufficiently high institutional longevity to give each policymaker confidence that the institution will persist during the period under which they are rewarded (or several subsequent periods on the iterated model).

The next three sections put some muscle on the proposed assembly's bones. The features that help the futures assemblies to satisfy the Epistemological Criterion and the Longevity Criterion will be woven throughout the following three sections, which focus on the assembly's job, powers, and tools; its structure; and its relationship with the rest of government, respectively.[6]

2 Duties, powers, and tools

At a very general level, the Futures Assembly is tasked with and incentivised to represent future generations. But providing the assembly with more specific, concrete duties will help to ensure that they are motivated, and have appropriate guidance, to pursue the right goals. What duties should national and international governments give to an assembly with the design features I've described, and with the functional role of representing the interests of future generations?

The two central duties of the office are to (i) exercise its powers and work with government to produce policy that promotes the welfare of future generations, and (ii) evaluate the decisions of one previous assembly and determine a reward for its members, as a function of their success in their duties. In the iterated retrospective accountability model, the reward is a function of the previous assembly's performance on (i) and (ii). On the standard retrospective accountability model, the reward is a function of the previous assembly's performance on only (i). As discussed, this reward may be a monetary reward for the previous assembly, or it may be a different sort of reward, such as a gift to their descendants, the continued promotions of their projects, or the protection of their legacy.

2.1 Policy duties

Let's consider the first duty of the assembly: to exercise its powers and work with the government to produce policy that promotes the welfare of future generations. Two parts of this duty require further elaboration and precisification. The first is the powers and tools that the assembly can use to achieve its goal. The second is what, precisely, it means for assemblies to have the goal of promoting the welfare of future generations.

A central set of tools that any futures assembly will have at its disposal are the government-facing 'soft powers' of persuasion, research and investigation, and advice-giving, as well as communication with the general public. More specifically, assemblies have the powers of a direct line to legislators, the ability to set their own research and policy agendas outside the reaches of the political business cycle, and the power to call on experts and convene

[6] An anonymous reviewer of this chapter noted that some political scientists (such as Achen and Bartels 2002; 2016) are sceptical of retrospective voting as it normally functions in democracy, in that they believe that voters are too ignorant to accurately apportion credit to previous administrations and are therefore swayed in their voting by entirely irrelevant, emotionally salient information (e.g. recent shark attacks). However, the most important 'finding' in this literature has failed to replicate (Fowler and Hall 2018), and the epistemic situation of members in the proposed assembly is in many ways superior to the epistemic situation of members of an electoral democracy. After all, retrospective voting is these members' *entire jobs* for multiple years (subsection 2.2), the members have much better access to information due to support from experts and archivists (subsections 3.1–3.3), and they are less polarised than ordinary members of the demos due to effects from sortition and deliberation (subsection 1.1).

summits. They also have high institutional legitimacy and status and a public relations (PR) team to share ideas and craft public narratives to inspire movements among the public.

Despite not having any teeth, even purely soft-power political institutions often have a very significant effect on government decision-making. Many citizens' assemblies have this profile, such as the aforementioned Irish Citizens' Assembly, which provoked a referendum to remove Ireland's constitutional ban on abortion and substantially shaped Ireland's Climate Action Plan. In-government research institutes such as the Office of Technology Assessment (OTA) have no formal powers at all. Yet a 1990 study by the Carnegie Commission on Science, Technology, and Government found that OTA reports were 'useful' to 'very useful' to 91% of congressional staff (Bimber 1990). One analysis found that the OTA's 1980s studies on synthetic fuels 'helped secure approximately $60 billion in savings' (Tudor and Warner 2019).

Recent research suggests that another soft-power institution, the European ombudsman, also has a significant effect on government policy (Finkel 2006; Koo and Ramirez 2009: 1330; Reif 2011: 286; Beckman and Uggla 2016), and even that it played a key role in the democratisation of European Communist countries (Gilligan 2010: 578). For example, Baranovsky (2016) finds that 'in a fixed effects model, both the length of service of the [Russian Ombudsman for Human Rights] and the relative busyness of the office (number of complaints per capita) have had an effect on a region's corruption index'.

The second set of powers to review the structure, staff, and rules of the futures assembly are similarly 'soft', in that they do not have the power to coerce or compel any other government institution. This set of powers is key to ensuring the relevance of the institution for many generations to come.[7]

These two powers of persuasion and of self-review make up the core tools that a soft-power futures assembly can use to promote its policy agenda. An institution with just these powers will likely do best by the lights of the Longevity Criterion. A major reason that future-oriented institutions are repealed today is that they have too much formal power (Jones, O'Brien, and Ryan 2018). So the future-oriented institution that is least likely to be repealed would plausibly have very minimal formal powers.[8]

[7] In particular, any promising futures assembly must be empowered to review the scoring rules that are used to determine the bonuses of the previous assemblies, along with the details of the rules about how to archive information and select experts. As each of these sets of rules has a powerful effect on the incentives of the institution, I suggest that the power of review can only be used in such cases if two successive assemblies achieve supermajority support for the proposed rule changes, and that the changes to the assembly's rules only go into effect when a third successive assembly comes into office. This system of review reduces groupthink, prevents assemblies from changing the rules to their own advantage, and ensures the assemblies are not entirely self-governing.

[8] That said, there are several reasons we might consider expanded powers for a futures assembly: it may be more feasible in the future to sustain a robust, future-oriented political institution, and the assembly may need expanded powers to fully achieve its policy goals or for bargaining or self-protection. What sort of 'harder powers' might futures assemblies ideally possess? Assemblies might, as a few examples, (i) require the legislature to read and respond to the assembly's reports and recommendations; (ii) require the government to notify the public about a proposed policy; (iii) require 'posterity impact assessments', or reports on the future social or economic impact of a policy proposal (John and MacAskill 2021); (iv) have the power to directly place a piece of proposed legislation in front of a legislative body and force a vote; (v) have the power to delay legislators from passing a potentially harmful piece of legislation (reviewable by judiciary or supermajority of legislatures, to prevent the assembly from obstructing a genuinely urgent policy need). With these quasi-legislative powers, the assembly might aim to precommit governmental bodies into taking future actions which will later seem difficult but high impact, to extend or eliminate budget windows, or to force policymakers to amend their implicit discount rate on utility (for an overview of why non-zero discount rates on utility are ethically unjustifiable, see Greaves (2017)).

We've considered the first part of the futures assembly's duty to exercise its powers and work with the government to produce policy that promotes the welfare of future generations: the powers and tools that the assembly can use to achieve its goal. Now we can consider the second part: what, precisely, it means for assemblies to have the duty of promoting the welfare of future generations. Here are three key questions: (i) What normative framework should assemblies apply to evaluate welfare changes in a generation? (ii) On what time horizon should the assemblies operate? (iii) What issues should the assemblies prioritise?

The first question is perhaps the most consequential for what the assembly's goals should be, but it is not one that I can answer here. What particular philosophical theory of well-being futures assemblies should adopt, and the broader question of how they rank states of affairs with respect to value, is, in modern democracies, a matter for public deliberation, and must ultimately be beholden to the deliberations of each assembly, to allow for these assemblies' self-government and for changes in the citizens' terminal goals as our moral understanding advances and our preferences change.

It is tempting to answer the second question in a similar way. However, there are several reasons why the rough time horizons of the futures assemblies should be determined by prior legislation. The first is that the time horizons of futures assemblies must be consonant with their institutional incentives. If, for example, an assembly is required to promote the interests of people more than one century from now, but is rewarded 30 years into the future, then it will be poorly incentivised to achieve its own mandates, and indeed will face the dualing, contradictory incentives of meritocratic reward and intrinsic motivation to act on the prescribed mandates. In addition to weakening their incentives to do any particular thing, this could cause confusion about goals and internal divisions within the assembly. The second reason is the need for temporal coordination across the various assemblies, to ensure that all generations are protected. To see how this could be a problem, imagine that the assemblies in power from 2030 to 2059 optimise for 30-year time horizons, whereas the assemblies in power from 2060 optimise for 60-year time horizons. In such a case, no futures assembly is responsible for protecting the people who live from the years 2090 to 2120. If assemblies all adopt the same time horizons, we can ensure the protection of all generations. With that said, our ability to predict the future changes with time. Sometimes forecasting becomes easier due to better tools and techniques, and sometimes it becomes more difficult due to entrance into periods of explosive growth and change. And if only for this reason, time horizons should be subject to adaptation through the futures assemblies' intentionally sluggish powers of self-review, alongside corresponding changes to the time horizons of the institutions' incentives mechanisms, if two successive assemblies agree to such changes.

On to the third question: What issues should assemblies prioritise? At a very basic level, it is a fairly straightforward matter that futures assemblies should consider all areas of policy that concern future generations on the relevant time horizon. This is because decisions in one policy domain that affects future generations (such as green energy) very often have effects in another policy domain that affects future generations (such as the workforce or economic growth). To have balanced policy that attends to all of the needs of future generations, and does not attend to one need at the expense of others, any sole institution charged

with representing future generations must consider all areas of policy that concern future generations.[9]

2.2 Evaluation duties

With the first set of duties precisified, and the powers to execute these duties elaborated, we can now turn to the assemblies' second set of duties: to evaluate the decisions of one previous assembly and determine a reward for its members, as a function of their success in their duties.

While I have saved the discussion of evaluation duties for last, these are the first duties that a futures assembly should work to fulfill. Futures assemblies are constituted by ordinary citizens, so they have a lot to learn before they can begin engaging in successful policymaking. The exercise of a futures assembly's evaluation duties are part of its early learning phase—a period of about one year wherein assembly members consult with experts on issues of long-term importance; evaluate the work of a previous assembly; and attend seminars about civics, the legal structure and history of their office, and policy matters of long-term importance. It is also in this period that the assembly begins to set policy goals for their remaining time in office.

The duty to evaluate predecessors is an essential part of the retrospective accountability mechanism that incentivises assembly members to act for the long term. But it is also an essential part of the learning and socialisation process required by the Epistemological Criterion. The assemblies will learn much from the process of working to understand the successes, mistakes, and lessons learned by the previous generation. Moreover, the scoring rules that the assemblies use to evaluate the previous generation give them a score card against which they can calibrate their own decisions, and see what choices of the previous generation were helpful or unhelpful to achieving a high score.

The score card that an assembly applies to the previous generation, and then uses to reflect on its own actions, is made up of a combination of high-level and low-level scoring rules. The high-level scoring rules are nigh-immutable goals that are sufficiently general to be applicable to every futures assembly, regardless of how the world changes. These scoring rules, then, are exactly the rules set out as duties in the previous section. Each assembly needs to ask of the previous generation: How well did the assembly fulfill its duty to exercise its powers and work with the government to produce policy that promotes the welfare of future generations? On the iterated variant, the assembly must also ask: How well did the assembly fulfill its duty to evaluate the decisions of one previous assembly and determine a reward for its members, as a function of their success in their duties? (This in turn will require examining the assembly two generations prior, to evaluate the accuracy of the previous assembly's judgments.)

The low-level scoring rules are malleable and set by the assemblies themselves. These can contain any number of objectives, including: (i) high-level, abstract goals such as peace

[9] It might be too epistemically demanding for any small group of citizens to make informed, sensible decisions across all domains. One tentative solution is to divide the intellectual labour *temporally*—e.g. giving the first assembly a generalist focus, which can in turn determine a particularly pressing policy priority for the second assembly to specialise in, then rotating back to a generalist focus for the third assembly, and so on.

and security; (ii) concrete, long-term metrics such as annual increase in GDP, target population levels, annual decrease in national Gini index, target levels of unemployment, target levels of atmospheric CO2 levels, or target levels of existential risk as reported in prediction markets; and (iii) concrete, short-term goals and metrics such as international agreement to an OECD beneficial artificial intelligence (AI) treaty, or the target size of the global nuclear arsenal. Each assembly inherits the low-level scoring rules from the assembly holding office just before it and has the flexibility to refine, change, or discard these rules in any way it pleases, as part of the learning phase.

The previous generation should be scored in part based on their success at achieving their own low-level scoring rules, but they should also be scored on their success at identifying the best low-level scoring rules for their assembly, creating an incentive for each assembly not only to act consistently with their scoring rules, but also to adopt ambitious and appropriate scoring rules for themselves.[10]

3 Structure

This section spells out one promising structure for the futures assembly. First is the machinery of the institution that helps it carry out its various duties: its archival arm, relationships with experts, research and investigation arm, and media and public relations team. Second is the description of the deliberating body of citizens itself: its constitution, size, salary, and term length.

3.1 Archivists

The archivists are the assembly's 'institutional memory': an independent bureaucratic body with no political status or access to the public, paid to (i) record all of the decisions and deliberations made by the assembly, (ii) package reports, evaluations, conference proceedings, and other materials for later assemblies, and (iii) chronicle the long-term social, economic, environmental, and political trends that are identified by the assembly as worthy of ongoing observation.

The sole purpose of the archivist arm is to help solve epistemological obstacles faced by futures assemblies. By recording decisions and deliberations and packaging all of the ideas of the assembly in a way that makes them easy to absorb by later assemblies, they significantly reduce the epistemic burdens of retrospective accountability. By chronicling long-term trends, at the instruction of futures assemblies, they help solve 'creeping problems'—the neglect of long-term issues, such as environmental issues, that happen as

[10] One crucial question is whether assemblies should evaluate previous generations from an *ex post* perspective—how good were the assembly's decisions, given how the world turned out?—or an *ex ante* perspective—how good were the assembly's decisions, given the information available to them at the time? An *ex ante* system seems more epistemically tractable for the previous generation, whereas an *ex post* system seems more epistemically tractable to the evaluating generation. Neither seems unqualifiedly preferable to the other. I suggest that whichever system is adopted properly balances the epistemic tractability burdens between both generations, and properly incentivises the prior generation to seek out new information.

a consequence of an aggregate of many small, soritical changes over long time horizons (Glantz 1999).

3.2 Experts

The experts serve to provide background seminars on a wide variety of issues relevant to assessing long-term priority areas and the long-term effects of policy. They also serve as consultants on the decisions and deliberations of the futures assembly. Experts should come from every discipline relevant to future-oriented policy, including the natural sciences, the social sciences, and philosophy, as well as relevant interdisciplinary endeavours such as futures studies, forecasting, and foresight. Their sole function is to infuse technocratic expertise into the deliberations of the assembly.

Experts are generally academics, researchers, or public intellectuals with significant academic training. They are paid a stipend for their contributions and may be bought out from various teaching and research requirements, but they are not expected to support the futures assembly in a full-time capacity. Their term length can be the same as the term length of the futures assembly, or longer.

3.3 Research and investigation arm

In addition to experts, the assembly also needs a research and investigation arm: a body of civil servants paid to carry out research directly commissioned by the futures assembly. This allows the assembly to expand its research and investigation capacity without requiring too much time from issue-area experts. The research and investigation arm should be as large as the futures assembly can productively make use of, potentially with many more members than the assembly itself. Each futures assembly may have substantial control over who is hired and fired in the research and investigation arm, but it should have an operations staff with little turnover, which can assist with the management and hiring of researchers to reduce the time costs of doing so to the futures assembly.

3.4 Public relations

It is important to ensure that the futures assembly is well understood by and has strong communication with the general public, since institutional complexity is one of the most powerful determinants of public and elite refusal to invest in long-term policy priorities (Jacobs and Matthews 2012). Accordingly, all investigations, recommendations, meetings, and reports should be made available to the public and advertised widely by an internal public relations and media team. A primary goal of this project should be institutional transparency, with a secondary goal to craft compelling public narratives for the sake of movement-building and civic education. To ensure that the secondary goal does not become dominant, turning the PR team into a propaganda machine, among the public-facing documents should be the meeting notes taken by the independent archivists, which can be used to ensure the veracity of the PR team's messaging.

3.5 The assembly

The most promising futures assemblies would be relatively large—c. 100–600 members, depending on the population of the jurisdiction in which it is implemented—to ensure demographic representativeness and resist corruption from interest groups. To further guard against corruption and ensure representativeness and minimal resignations, assembly members should be paid a high salary, for example commensurate with the typical salary for members of the national legislative body, plus the bonus already described, which should similarly be large enough to appropriately motivate assembly members.

Futures assembly members serve single terms and cannot be re-appointed, to avoid short-termist re-election incentives. The complexity of the job suggests a fairly long term length, to allow adequate time for learning and policymaking. On the other hand, the assembly members are not professional politicians and are selected from throughout the country, so the terms need to be minimally disruptive for randomly chosen citizens who have other jobs and projects, suggesting a shorter term limit. Shorter term limits also restrict the extent to which assembly members can be reached by corrupting interest groups and allow more citizens to participate in the process over the long run. Nonetheless, a term length shorter than 3 years would not be adequate to the task of policymaking, so I suggest that a term length of 3–5 years is appropriate.[11] The potential disruptiveness of this work should be ameliorated with opportunities to perform most work remotely and with generous family support policies.

4 Checks and balances

How will this new institution interact with the rest of the government? In light of its powers, tools, and policymaking role, the futures assembly will primarily engage with the legislature, regulatory executive agencies, and other public bodies, to precipitate policy changes at every level of government organisation. While the purpose of the futures assembly is to influence the rest of government, however, special attention should be given to ensure that the futures assembly is not unduly influenced by the present, short-termist government, especially the legislature. This is essential so that the futures assembly can insulate itself from the political business cycle and make reflective decisions about long-term matters outside of the news cycle and public attention. It is for this reason that the futures assembly is empowered to set its own research and policy agenda, and has its own research staff to investigate matters of long-term importance.

The power of the futures assembly must of course be checked by the rest of the government so that it does not become excessively powerful. Fortunately this is no great risk, given

[11] An anonymous referee suggests that the need for selected individuals to learn everything relevant to their job counts in favour of asynchronous terms, so that at any given time there are new members and more experienced members who can teach the new members. This is a plausible suggestion, but it also makes it more likely that the futures assembly develops a distinctive internal culture that is transmitted from cohort to cohort, which could make its culture deviate increasingly from the surrounding political culture. This may make the assembly's epistemic and motivational features more locked in and less susceptible to change with novel cohorts. It is not clear these features would affect the success of the scheme overall.

that almost all of the powers proposed for the futures assembly herein have been advisory powers, and given that the futures assembly can at any time be repealed by the same legislative action that would give birth to it in the first place.[12] The only substantial, binding power proposed for the futures assembly is the power to delay legislation (see fn. 7), and this is to be checked either by the judiciary or by a supermajority of the legislature, as indicated previously.

5 Conclusions

Political short-termism costs the global economy many billions and perhaps trillions of dollars annually and leads to many millions of deaths from disasters and suboptimal resource allocation. In this chapter, I proposed that a futures assembly explicitly incentivised to promote the interests of future generations might be a promising strategy to ameliorate political short-termism. The most initially promising such retrospective accountability mechanism is an iterated mechanism, in which each generation of policymakers rewards the previous generation for their policy choices and for their own evaluation of the previous generation.

The representative mechanism proposed is among the most advanced and promising in the literature. But its potential for success is plagued with uncertainty. For one, while the mechanism is supported by contemporary political science and armchair theory, nothing like it has ever been implemented, and it is unclear whether the model's assumptions about political motivation will hold up in practice. Second, I have not yet discussed the implementability of this reform. Given the widespread use of citizens' assemblies and the creation of several future-focused institutions in the last two decades, one should not assume that the proposal is nowhere feasible. But given that few *permanent* citizens' assemblies exist, and that the accountability mechanism is novel, it is likely to be somewhat less implementable than other recent reforms.[13]

Most importantly, while the mechanism may incentivise policymakers to act with the next hundred to a thousand years in view, it is very unlikely to be able to incentivise policymakers to act with the next million years in view. This is a problem, since very plausibly governments ought not be biased towards the near-term *at all*, and since vastly more people face the consequences of our actions over the next million years than will over the next thousand years, in expectation.

There is some reason to think that the interests of future generations over the next thousand years are highly correlated with the interests of future generations in a million years, since both are affected by the quality of intergenerational global public goods such

[12] Given the possibility of repeal, the relevant jurisdiction should adopt measures to ensure that the futures assembly will still be evaluated and paid by an ad hoc body of auditors at the appropriate time, even though the institution has itself been repealed. This will increase every assembly's confidence in the persistence of the incentives mechanism, increasing their willingness to pay personal costs for the sake of future generations, given the higher guarantee of a reward for doing so.

[13] Some political theorists have discussed this issue under the label 'the bootstrapping problem'. As the argument sometimes goes, there is something odd or surprising about *any* proposed reform that asks citizens to vote for a policy that then binds them to take more action than they would otherwise be willing to do. But these political theorists overestimate the challenge. It is a widespread feature of tax policy (e.g. redistributive policy) and consumer policy (e.g. cage-free egg policy) that citizens who vote for these policies do so even though they would not voluntarily increase their spending on government programmes or voluntarily reduce their consumption of goods with negative externalities. However puzzling this is from the armchair, it is a well-established empirical fact.

as the environment, resources, technological progress, knowledge acquisition, and human survival—and in these domains, I think this much is true. But to the extent that the interests of near-future and far-future generations are decorrelated, it is very unlikely that the proposal succeeds in promoting the well-being of far-future generations. Future work could improve on the model proposed by extending the time horizon of political decision-making even further, or by developing more narrowly-targeted mechanisms that incentivise policy-makers to act specifically on policy objectives that benefit near- and far-future generations, such as permanently civilisation-ending catastrophes.

Acknowledgement

This chapter was greatly improved by feedback from Greg Bognar, Mark Budolfson, Adam Gibbons, Axel Gosseries, Alex Guerrero, Larry Temkin, and one anonymous referee.

Bibliography

Achen, C. H. and Bartels, L. M. (2002), 'Blind Retrospection: Electoral Responses to Drought, Flu, and Shark Attacks', presentation (Annual meeting of the American Political Science Association, Boston).

Achen, C. H. and Bartels, L. M. (2016), *Democracy for Realists: Why Elections Do Not Produce Responsive Government* (Princeton University Press).

Aoki, R. and Vaithianathan, R. (2012), 'Intergenerational Voter Preference Survey – Preliminary Results' (CIS Discussion Paper Series).

Arnold, R. D. (1990), *The Logic of Congressional Action* (Yale University Press).

Banks, J. S. and Sundaram, R. K. (1993), 'Adverse Selection and Moral Hazard in a Repeated Elections Model', in W. A. Barnett, M. J. Hinich, and N. J. Schofield (eds.), *Political Economy: Institutions, Competition, and Representation* (Cambridge University Press), 295–311.

Barnovsky, A. (2016), 'The Ombudsman Effect: How Effective is the Institution of the Human Rights Commissioner in Russia's Regions?', working paper, https://scholar.harvard.edu/files/alla-baranovsky/files/ombuds.paper_.pdf (accessed 30 November, 2021).

Beck, N. (1982), 'Does There Exist a Political Business Cycle: A Box-Tiao Analysis', in *Public Choice* 38: 205–209.

Beckman, L. and Uggla, F. (2016), 'An Ombudsman for Future Generations', in I. González-Ricoy and A. Gosseries (eds.) *Institutions for Future Generations* (Oxford University Press), 117–134.

Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far future*. Doctoral thesis (Rutgers University).

Bimber, B. (1996), *The Politics of Expertise in Congress: The rise and fall of the Office of Technology Assessment* (The State University of New York Press).

Binder, S. A. (2006), 'Can Congress Legislate for the Future?', Research Brief 3 (John Brademas Center for the Study of Congress, New York University).

Bohman, J. (2007), *Democracy Across Borders: From Dêmos to Dêmoi* (MIT Press).

Brennan, J. (2017), *Against Democracy* (Princeton University Press).

Brighouse, H. and Fleurbaey, M. (2010), 'Proportionality and Democracy', in *Journal of Political Philosophy* 18: 137–155.

Caney, S. (2016), 'Political Institutions for the Future: A Five-Fold Package', in I. González-Ricoy and A, Gosseries (eds.) *Institutions for future generations* (Oxford University Press), 135–155.

Climate Assembly UK. (2020), *The Path to Net Zero: Climate Assembly UK Full Report*, https://www.climateassembly.uk/report/read/final-report.pdf (accessed 15 January 2025).

Coleman, M., Devaney, L., Torney, D., and Brereton, P. (2019), 'Ireland's World-Leading Citizens' Climate Assembly. What Worked? What Didn't?' *Climate Home News*, https://www.climatechangenews.com/2019/06/27/irelands-world-leading-citizens-climate-assembly-worked-didnt/ (accessed 15 January 2025)

Elster, J. (2008), *Alchemies of the Mind: Rationality and the Emotions* (Cambridge University Press).

Fearon, J. (1999), 'Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance', in A. Przeworski, S. C. Stokes, and B. Manin (eds.) *Democracy, Accountability, and Representation* (Cambridge University Press): 55–97.

Ferejohn, J. (1986), 'Incumbent Performance and Electoral Control', in *Public Choice* 50: 5–25.

Finkel, E. (2006), 'Defending Rights, Promoting Democracy', working paper (The Institution of Ombudsman in Poland, Russia and Bulgaria).

Fiorina, M. P. (1977), 'An Outline for a Model of Party Choice', in *American Journal of Political Science* 21: 601–625.

Fishkin, J. S. and Luskin, R. C. (2005), 'Experimenting with a Democratic Ideal: Deliberative Polling and Public Opinion', *Acta Politica* 40: 284–298.

Fishkin, J. S., Mayega, R. W., Atuyambe, L., Tumuhamye, N., Ssentongo, J., Siu, A., and Bazeyo, W. (2017), 'Applying Deliberative Democracy in Africa: Uganda's First Deliberative Polls', *Daedalus* 146: 140–154.

Fowler, A. and Hall, A. B. (2018), 'Do Shark Attacks Influence Presidential Elections? Reassessing a Prominent Finding on Voter Competence', in *The Journal of Politics* 80/4: 1423–1437.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002), 'Time Discounting and Time Preference: A Critical Review', in *Journal of Economic Literature* 40: 351–401.

Gilligan, E. (2010), 'The Human Rights Ombudsman in Russia: The Evolution of Horizontal Accountability', in *Human Rights Quarterly* 32/3: 575–600.

Glantz, M. H. (1999), *Creeping Environmental Problems and Sustainable Development in the Aral Sea Basin* (Cambridge University Press).

González-Ricoy, I. and Gosseries, A. (2016), 'Designing Institutions for Future Generations: An introduction', in I. González-Ricoy and A. Gosseries (eds.) *Institutions for Future Generations* (Oxford University Press), 3–23.

Goodin, R. E. (1995), *Utilitarianism as a Public Philosophy* (Cambridge University Press).

Greaves, H. (2017), 'Discounting for Public Policy: A Survey', in *Economics and Philosophy* 33/3: 391.

Greaves, H. and MacAskill, W. (2021), 'The Case for Strong Longtermism', GPI Working Paper No. 5-2021 (Global Priorities Institute, Oxford University), https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/ (accessed 30 November 2021).

Grönlund, K., Herne, K., and Setälä, M. (2017), 'Empathy in a Citizen Deliberation Experiment', in *Scandinavian Political Studies* 40: 457–480.

Healy, A. and Malhotra, N. (2009), 'Myopic Voters and Natural Disaster Policy', in *American Political Science Review* 103: 387–406.

Jacobs, A. M. and Matthews, J. S. (2012), 'Why Do Citizens Discount the Future? Public Opinion and the Timing of Policy Consequences', in *British Journal of Political Science* 42: 903–935.

John, T. M. (2020), 'Longtermist Institutional Design and Policy: A Literature Review' (unpublished manuscript), https://docs.google.com/document/d/1DdKjaRG0OOu7M1MB40bgdlJSykENQdoD8q28Yw9nuuc/ (accessed 15 January 2025)

John, T. M. (2022), *Chronopolitanism: Political Inclusion and the Distant Future*, PhD thesis, Rutgers University–New Brunswick.

John, T. M. (2023), 'Empowering Future Generations by Empowering the Young?', in G. Bognar and A. Gosseries (eds.) *Ageing without Ageism? Conceptual Puzzles and Policy Proposals* (Oxford University Press), 143–158.

John, T. M. and MacAskill, W. (2021), 'Longtermist Institutional Reform', in T. M. John and N. Cargill (eds.), *The Long View* (FIRST), 45–60.

Johnson, D. and Levin, S. (2009), 'The Tragedy of Cognition: Psychological Biases and Environmental Inaction', in *Current Science* 97: 1593–1603.

Jones, N., O'Brien, M., and Ryan, T. (2018), 'Representation of Future Generations in United Kingdom Policy-Making', in *Futures* 102: 153–163.

Kamijo, Y., Teruyuki, T., and Yoichi, H. (2020), 'Effect of Proxy Voting for Children under the Voting Age on Parental Altruism towards Future Generations', in *Futures* 122: 102569.

Key, V. O. Jr. (1966), *The Responsible Electorate: Rationality in Presidential Voting 1936–1960* (Harvard University Press).

Koo, J. and Ramirez, F. O. (2009), 'National Incorporation of Global Human Rights: Worldwide Expansion of National Human Rights Institutions, 1966-2004', in *Social Forces* 87/3: 1321–1353.

List, C., Luskin, R. C., Fishkin, J. S., and McLean, I. (2013), 'Deliberation, Single-Peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls', in *The Journal of Politics* 75: 80–95.

MacKenzie, M. K. (2016), 'Institutional Design and Sources of Short-Termism', in I. González-Ricoy and A. Gosseries (eds.) *Institutions for Future Generations* (Oxford University Press), 24–46.

Mayhew, D. R. (1974), *Congress: The Electoral Connection* (Yale University Press).
Norwood, F. B., Tonsor, G., and Lusk, J. L. (2019), 'I Will Give You My Vote but Not My Money: Preferences for Public versus Private Action in Addressing Social Issues', in *Applied Economic Perspectives and Policy* 41: 96–132.
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Hachette).
Paul, A. S., Lusk, J. L., Norwood, F. B., and Tonsor, G. T. (2019), 'An Experiment on the Vote-Buy Gap with Application to Cage-Free Eggs', in *Journal of Behavioral and Experimental Economics* 79: 102–109.
Reif, L. C. (2011), 'Transplantation and Adaptation: The Evolution of the Human Rights Ombudsman', in *Boston College Third World Law Journal* 31: 269–310.
Rogoff, K. (1990), 'Equilibrium Political Budget Cycles', in *American Economic Review*, 80: 21–36.
Spitzer, R. J. (1988), *The Presidential Veto* (SUNY Press).
Steiner, J., Bächtiger, A., Spörndli, M., and Steenbergen, M. R. (2005), *Deliberative Politics in Action: Analyzing Parliamentary Discourse* (Cambridge University Press).
Tarsney, C. J. (2019), 'The Epistemic Challenge to Longtermism'. GPI working paper (Global Priorities Institute, University of Oxford), https://globalprioritiesinstitute.org/wp-content/uploads/2019/Tarsney_Epistemic_Challenge_to_Longtermism.pdf (accessed 15 January 2025).
Warner, J. and Tudor, G. (2019), 'The Congressional Futures Office.' (Belfer Center for Science and International Affairs, Harvard Kennedy School).
Tufte, E. R. (1978), *Political Control of the Economy* (Princeton University Press).
Weber, E. U. (2006), 'Experience-Based and Description-Based Perceptions of Long-Term Risk: Why Global Warming Does Not Scare Us (Yet)', in *Climatic Change* 77: 103–120.

28

Longtermism and Social Risk-Taking

H. Orri Stefánsson

1 Introduction

Some risks seem unacceptable when considered on their own, even though they seem acceptable when considered as part of a larger bundle of risks. For instance, while those who are risk averse with respect to money might turn down a 50/50 gamble between losing $100 and winning $200, few people would turn down a bundle of 100 such independent[1] gambles (see, e.g., Samuelson 1963; Rabin 2000; Kahneman 2012). After all, such a bundle has a monetary expectation of $5,000 and has only a 0.04% chance of resulting in monetary loss. As Rabin and Thaler (2001: 223) put it: 'A good lawyer could have you declared legally insane for turning down this [bundle].'

Something similar would seem plausible when taking risks that affect others. For instance, suppose that a physician is considering a risky intervention that has a half chance of costing a patient 1 unit of wellbeing and a half chance of benefitting the patient by 2 units of wellbeing. Then if losing 1 unit of wellbeing is significant, the physician might reasonably choose not to make the intervention; in fact, they will not make the intervention if they are moderately risk averse and/or moderately loss averse with respect to the wellbeing of others, in a sense to be made precise in the next section.

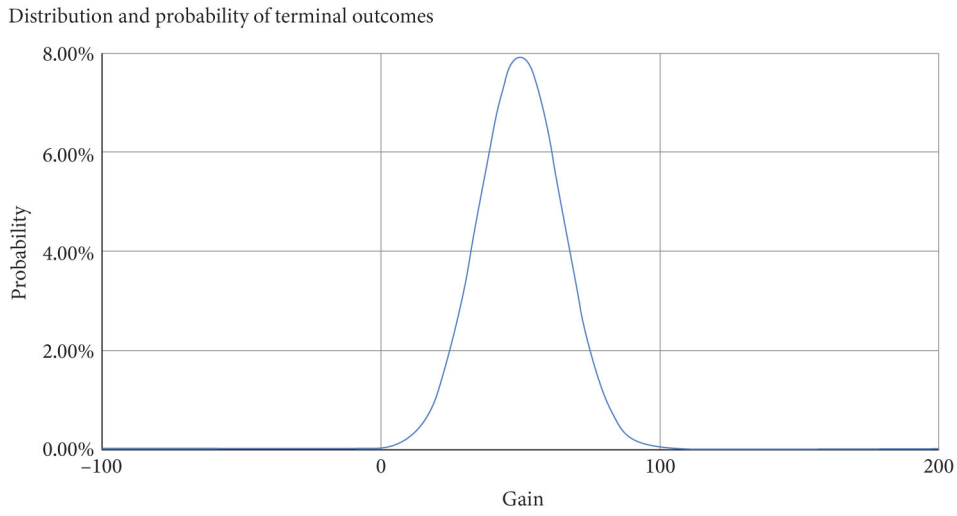
But now suppose instead that the intervention affects 100 patients, where for each patient the intervention is a 50/50 gamble between losing 1 unit of wellbeing and gaining 2 units, and the gambles are probabilistically independent. In that case, it seems that even a moderately risk or loss averse decision-maker should choose the intervention.[2] After all, the expected total wellbeing effect of the bundle is a gain of 50 units, and the probability that the bundle results in a total wellbeing loss is minuscule (0.0004).[3] In addition, the bundle can at most result in each person losing 1 unit of wellbeing, and each patient's gamble provides them with positive expected wellbeing. The figure shows the distribution and probability of total wellbeing from the bundle.

[1] Here and elsewhere, gambles are 'independent' if the probability distribution associated with any one of the gambles is independent of the outcome of any of the other gambles.

[2] It might be worth noting that some would not accept the bundle, for instance, those who accept some version of *leximin*, and thus always choose the gamble whose worst possible result is at least as good as the worst possible result that any other gamble might resulting in. Such a rule is arguably not *moderately* risk averse. The same is true of some *ex post* contractualists.

[3] In fact, the probability of losing more than 50 units in total is merely 0.0000000000013! In contrast, the probability of gaining more than 50 units is 0.46.

H. Orri Stefánsson, *Longtermism and Social Risk-Taking* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © H. Orri Stefánsson 2025.
DOI: 10.1093/9780191979972.003.0028



More generally, the more gambles that a risk or loss averse 'social planner' considers together, the more prone they should be to accept each gamble, assuming for instance that each gamble has a positive expectation in whatever *objective* quantity the planner takes to be the object of risk-free axiology (for instance, wellbeing or years in full health).

Now, it seems plausible that to truly internalise the long-term perspective (or 'longtermism'; more on this in a moment) means to judge gambles in light of what one predicts about the (long-term) future. Furthermore, some think that one should treat a sequence of gambles as one would treat a bundle of those gambles. But then the above argument seems to show that taking the long-term perspective should often make a social planner more risk prone, as compared with a 'short-termist' social planner with the exact same attitudes to risks, gains, and losses. In the next section I present a more formal version of this argument.

But first, let's consider a rather different and, in some ways, less abstract argument for the same conclusion. Suppose that a social planner is considering a risky social change—an 'experiment'—such as legalising recreational drugs. The social planner thinks that, in terms of aggregate wellbeing, the experiment has a positive expectation; let's say they take this to be a 50/50 gamble between the population of interest losing in total 100 units of wellbeing per generation and gaining in total 200 units of wellbeing per generation. Again, if the social planner is risk or loss averse with respect to the population's total wellbeing, then they might not be willing to take the risk if they focus on the effect on the current and perhaps the next few generations, since they might judge that they cannot justify exposing them to a half chance of losing 100 units of wellbeing for the sake of the same chance of gaining 200 units. However, if they take a longer perspective, then they might reach a different conclusion. As long as a harmful experiment can be stopped, they might reason that the long-term benefit of a gain of 200 units of wellbeing per generation if the experiment is a success, in addition to the knowledge that the experiment is expected to bring—and assuming that that knowledge will be put to good use[4]—would outweigh the risk of harm to the current

[4] For an illuminating discussion of this and related issues, see Barrett and Buchanan (2023), who point out that an argument for what they call 'progressive experimentalism' has to address the *uptake problem*, that is, the possibility that the knowledge gained from a social experiment will not be put to good use.

generation. Indeed, as formally demonstrated in the next section, this is precisely how a moderately risk or loss averse social planner should reason.

The main objective of this chapter is to carefully formulate two arguments that what I will call 'longtermism' should make a risk or loss averse social planner more risk prone; and, in particular, to consider the conditions needed for these arguments. By 'longtermism' I shall simply mean the claim that a social planner should take the long-term perspective. This is, of course, admittedly quite vague (and, in fact, I shall discuss two distinct notions of 'taking the long-term perspective', corresponding to the aforementioned two arguments). However, as will become evident, the precise meaning of 'longtermism' (e.g., *how* long and 'wide' a perspective the longtermist takes) is unimportant for my argument; a small shift in perspective beyond the present moment is in fact enough for my argument.

It might be worth acknowledging that what I am calling 'longtermism' isn't quite what Greaves and MacAskill (2021) call (axiological strong) 'longtermism': informally, the claim that 'far-future effects are the most important determinant of the value of our options' (Greaves and MacAskill 2021: 3). However, I think that longtermism as they understand it implies what I am calling 'longtermism'. If 'far-future effects are the most important determinant of the value of our options', then it would seem we *should* take the (very) long-term perspective. For if we do not, then we will simply ignore the most important determinants of the value of our options.

In the next section, I spell out two formal arguments for the claim that (what I call) longtermism supports social risk-taking. An important part of my analysis will consist in examining the conditions required for the success of these arguments. These conditions will also be made precise in the next section. In the third and final section, I consider to what extent these conditions hold for real-life risky public policies.

2 The formal arguments

2.1 Framework and definitions

To keep the argument as simple as possible, I will assume throughout this chapter some version of *generalised utilitarianism* and a population of a fixed size. An example of generalised utilitarianism is, of course, standard utilitarianism, according to which the value of a population is determined by the sum total of wellbeing in the population; another example is prioritarianism, according to which the value of a population is determined by the sum of priority-weighted wellbeing in the population. In what follows, I shall use standard utilitarianism for illustrative purposes, but my argument can be easily generalised to, say, prioritarianism.

Let $\mathbf{O} = \{o_1, \ldots, o_m\}$ be the set of m (terminal, i.e., risk-free) outcomes. According to generalised utilitarianism, we can take each outcome to be a vector of wellbeing levels, e.g., $o_i = (z_i^1, \ldots, z_i^n)$ given a population of n people. Gambles (or 'lotteries') are probability[5] distributions over $\mathbf{O}$. Let $\mathbf{L}$ be the set of gambles, that is, the set of all vectors $(\lambda_1, o_1; \ldots; \lambda_m, o_m)$ such that $\lambda_1, \ldots, \lambda_m \in [0,1]$ and $\sum_{k=1}^{m} \lambda_k = 1$. Let $\succsim$ be a (weak) better-than relation (or social

[5] I shall not make any particular assumptions about how or why these probabilities are available to the decision-maker. But it may be worth noting that the arguments I consider can be made for subjective as well as objective probabilities, and they can be extended to imprecise probabilities.

preference relation) defined on both **O** and **L**. The expression '$o_i \succsim o_j$' should be read as saying that outcome o_i is a least as good as outcome o_j. Correspondingly, '$o_i \sim o_j$' means that o_i and o_j are equally good, and '$o_i \succ o_j$' that o_i is strictly better than o_j. Finally, let u be a real valued utility function on the set of outcomes (unique up to positive affine transformations[6]) that represents $\succsim$ on **O** in the sense that $o_i \succsim o_j \Leftrightarrow u(o_i) \geq u(o_j)$.[7]

By risk aversion, I do *not* mean that the quantities of interest are taken to have decreasing marginal value, that is, I do not mean what economists typically mean by risk aversion. Elsewhere, I have argued that that picture fails to capture important aspects of risk aversion (Stefánsson and Bradley 2019). Informally, by risk aversion, I mean that the value that any of the *better* potential outcomes from a gamble contributes towards the overall value of the gamble is less than the probability-weighted utility of that outcome (Buchak 2013; Stefánsson and Bradley 2015; 2019). The version of this idea that is best-known amongst philosophers is due to Buchak (2013). Hers is also a particularly tractable version of this idea, which I will hence use below, for illustrative purposes.

We can now state different theories about how to manage risk and choose between gambles, before defining risk aversion more formally. According to expected utility theory, understood as a theory of rationality, one's preferences between gambles should correspond to how the gambles' expectations of utility compare. A gamble's expectation of utility is found by first weighing the utility of each possible outcome of the gamble by its probability, and then adding together these probability-weighted utilities. More formally:[8]

Expected utility (EU) theory. *For any gambles* $L_\alpha = (\alpha_1, o_1; \ldots; \alpha_m, o_m), L_\beta = (\beta_1, o_1; \ldots; \beta_m, o_m) \in L$ *and for any rational* $\succsim$:

$$L_\alpha \succsim L_\beta \Leftrightarrow \sum_{i=1}^{m} u(o_i) \cdot \alpha_i \geq \sum_{i=1}^{m} u(o_i) \cdot \beta_i$$

Expected utility theory does not allow for risk aversion in the above sense. EU theory does allow for a particular kind of risk aversion: for instance, if the outcomes are quantities of money, then EU theory can accommodate risk aversion with respect to money in the sense that it allows that a decision-maker strictly prefers any sure sum of money to a lottery whose expectation is that same sum of money. This is achieved by stipulating a *concave* utility function, u. But since the expected utility formula, as stated above, implies that the value that *any* outcome contributes to a gamble's value equals the outcome's

[6] Saying that u is unique up to positive affine transformations means that if both u and u' represent $\succsim$, then there exist real numbers a and $b > 0$ such that $u' = a + u \cdot b$.

[7] A utility function is simply a numerical representation of a ranking. Therefore, if the ranking is based on a (morally) 'better-than' relation, then the corresponding utility function will be a (moral) value function. For instance, below we interpret the 'utility' of an outcome as the sum of wellbeing in the outcome.

[8] Note that, formally, the two gambles contain *the same* outcomes. But, for either gamble, many of the outcomes may receive a zero probability and thus not be *possible* outcomes from the gamble.

probability-weighted utility, expected utility theory cannot accommodate risk aversion with respect to utilities, which is what I mean by 'risk aversion' (for a discussion, see, e.g., Stefánsson and Bradley 2019).

By contrast, rank-dependent utility theory, defended for instance by Buchak (2013) (under the name of 'risk weighted expected utility theory'), allows for what I am calling risk aversion. Informally, rank-dependent utility theory evaluates a gamble by adding to the utility of the gamble's *worst* possible outcome a weighted version of the expected utility that the gamble offers over and above its worst possible outcome, where the weight in question depends on the relevant agent's attitude to risk. If the agent is *maximally* risk averse, then the weight in question turns the expectation into zero, such that a gamble is evaluated by its worst possible outcome. If the agent in question is an expected utility maximiser, however, then the weight leaves the expectation unaltered, meaning that the gamble is evaluated by its expected utility. In most applications, the weight is somewhere between these, that is, the agent is assumed to be risk averse but not maximally so.

To state this theory more formally, now let $\mathbf{O}_{[]} = \{o_{[1]}, \ldots, o_{[m]}\}$ be a non-decreasing re-ordering of $\mathbf{O}$ according to the preference relation of interest, meaning that for any i, $o_{[i+1]} \succsim o_{[i]}$. And let r be a real valued and increasing 'risk function' on probabilities, satisfying the constraint that $r(0) = 0, r(1) = 1$.

Rank-dependent utility (RDU) theory. *For any rational* $\succsim$ *there is an* r *such that for any gambles* $L_\alpha = (\alpha_1, o_1; \ldots; \alpha_m, o_m), L_\beta = (\beta_1, o_1; \ldots; \beta_m, o_m) \in L$:

$$L_\alpha \succsim L_\beta \Leftrightarrow u(o_{[1]}) + \sum_{j=2}^{m}\left(\left[u(o_{[j]}) - u(o_{[j-1]})\right] r\left[\sum_{i=j}^{m}\alpha_i\right]\right)$$
$$\geq u(o_{[1]}) + \sum_{j=2}^{m}\left(\left[u(o_{[j]}) - u(o_{[j-1]})\right] r\left[\sum_{i=j}^{m}\beta_i\right]\right)$$

In RDU theory, risk aversion is captured by a *convex* risk function, which implies that for each gamble, the better potential outcomes get a lower weight than they do according to EU theory. Following Buchak (2013), I will use $r(x) = x^2$ as a canonical example of a risk function for a risk averse decision-maker. With this risk function, the RDU of a 50/50 gamble between losing 1 unit (of utility) and gaining 2 units is:

$$-1 + \left(\frac{1}{2}\right)^2 * 3 = -\frac{1}{4}$$

Next, I define 'risk aversion' more generally. The definition makes use of the concept of a 'mean preserving spread'.[9] Informally, L_β is a mean-preserving spread of L_α's utilities if the two gambles offer the same mean (i.e., expected) utility even though the

[9] For a historically important discussion of the connection between mean-preserving spread and risk aversion, see Rothschild and Stiglitz (1970).

probability density function (or probability mass function, for discrete outcomes) associated with the former is more spread, that is, assigns greater probability to more extreme values.

Risk aversion. *A preference relation, $\succsim$, is (generally)*[10] *risk averse if $L_\alpha \succ L_\beta$ whenever L_β is a mean-preserving spread of L_α's utilities.*

To illustrate the above definition, suppose that $u(o_1) = 0$, $u(o_2) = 5$, $u(o_3) = 10$. Then $(\frac{1}{3}, o_1; \frac{1}{3}, o_2; \frac{1}{3}, o_3)$ is a mean-preserving spread of the utilities of $(0, o_1; 1, o_2; 0, o_3)$: they offer the same expectation of utility but the probability mass function associated with the former assigns positive probability to more extreme utilities. So, someone who is generally risk averse would prefer the latter to the former. And, indeed, with $r(x) = x^2$, the rank-dependent utility of the latter is higher than the former: 5 compared to 2.778. To connect this to the informal statement of risk aversion I gave above—that is, the idea that the value that any of the better potential outcomes from a gamble contributes towards the overall value of the gamble is less than the probability-weighted value of that outcome—note that for the above risk-free 'gamble' to get a higher value than the risky one, it must be the case that for the risky gamble, at least one of o_2 and o_3 contribute a value to the gamble that is smaller than their probability-weighted utilities.[11]

Now let's turn to what I call 'loss aversion'.[12] Unlike, for instance, *prospect theory* (Kahneman and Tversky 1979), rank-dependent utility theory does not (by itself) allow for loss aversion. Nor, in fact, can any theory that has been proposed as a *normative* theory for managing risk account for loss aversion. But loss aversion does not seem obviously unreasonable, at least intuitively, when viewed from the perspective of a social planner. For instance, it is not obviously irrational—and some might even think that it is morally right—that a social planner is more concerned with avoiding a loss in the present generation's welfare than with securing a comparable increase in the present generation's welfare.

Informally, a social planner who is loss averse in aggregate wellbeing finds that it is worse when their population loses aggregate wellbeing of magnitude x than it is good when their population gains aggregate wellbeing of magnitude x. To make this more precise, let SQ_t be the 'status quo' at time t. Given the assumption of generalised utilitarianism, we can treat this as the wellbeing distribution that holds at time t. We now define a time-relative moral value function V_t—that is, one function for each time t—and we assume that each such function represents the better-than relation at the corresponding time, $\succsim_t$. Then we can define:

[10] Sometimes we may be interested in preference relations that are not generally risk averse but rather risk averse for say wellbeing levels within some ranges but risk seeking or risk neutral with respect to other ranges. Such complexities can, however, be set aside for now.

[11] This assumes that the value of any 'gamble' that results for sure in some outcome is equal to the value (in this case, the 'utility') of that outcome, which RDU as formulated above of course implies.

[12] It may be worth emphasising that a decision-maker can be *both* loss averse and risk averse (and, moreover, that some choices can be explained by either risk aversion or loss aversion). To keep things simple, in what follows, I shall, however, assume that the social planner of interest is *either* loss averse or risk averse.

Loss aversion. *A preference relation,* $\succsim$*, is (generally)*[13] *loss averse with respect to aggregate wellbeing if for any time t, for any triple of wellbeing vectors* $(z_i^1,\ldots,z_i^m)=o_i$, $(z_j^1,\ldots,z_j^m)=o_j$ *and* $(z_k^1,\ldots,z_k^m)=o_k$ *such that*

$$\left(\sum_{l=1}^{m} z_i^l - \sum_{l=1}^{m} z_j^l\right) = \left(\sum_{l=1}^{m} z_j^l - \sum_{l=1}^{m} z_k^l\right) > 0,$$

if $(z_j^1,\ldots,z_j^m)=SQ_t$ *then any* V_t *that represents* $\succsim_t$ *satisfies:*

$$V_t(o_i) - V_t(o_j) < V_t(o_j) - V_t(o_k)$$

Less formally, at any time *t* and given any magnitude of wellbeing *x*, it is worse, according to a loss averse relation, that society's aggregate wellbeing decreases by *x* than it is good that society's wellbeing increases by *x*.

Now, someone might object to the terminology, *loss aversion*, for the following reason.[14] A decision-maker who is loss averse, as I have defined it above, isn't strictly speaking particularly concerned with *a person losing* wellbeing; for instance, if one person loses wellbeing of magnitude *x* while another person gains wellbeing of magnitude $y > x$, then that may well count as an overall improvement, according to such a decision-maker, even if *y* is only marginally greater than *x*.[15] What they are averse to is *loss in total wellbeing* compared to some status quo. Moreover, some might think that loss aversion is only plausible if it concerns losses to persons. I disagree. Loss aversion understood as aversion to specific persons losing wellbeing (as opposed to aversion to the population losing aggregate wellbeing) would have extremely anti-egalitarian implications: it would sometimes imply that we should not make an intervention that would both increase aggregate wellbeing and increase inequality, for instance, by redistributing resources from those who are better off to those who are worse off. So, I think that the loss aversion, as I have defined it, is more normatively appealing.[16]

So, let's now consider what a loss averse outcome axiology might look like. We can start by zero-normalising V_t around SQ_t, that is, $V_t(SQ_t)=0$ for any time *t* (which of course requires re-normalising when the distribution changes). Next define the function:

$$\varnothing(x)=\begin{cases} x & \text{if } x \geq 0 \\ 3x & \text{if } x < 0 \end{cases}$$

[13] We could similarly define loss aversion with respect to some wellbeing levels (or differences), but general loss aversion suffices for our purposes.

[14] I thank Andreas Schmidt for making this point.

[15] If the planner is a loss averse prioritarian, then we add the qualification that the person losing *x* is better off than the person gaining *y*.

[16] Why not conclude from the above that loss aversion isn't appealing at all, rather than settling for loss aversion with respect to aggregate wellbeing, as I am suggesting? Don't we also need a *positive* argument in favour of loss aversion with respect to aggregate wellbeing, in addition to the argument against loss aversion at the level of individual wellbeing? (I thank a referee for raising these questions.) Offering anything like a conclusive positive argument of this sort is beyond the scope of this chapter. Instead, the following brief suggestion will have to suffice. In general, it may seem plausible that one should be especially *cautious* when taking risks on other people's behalf, in particular, when one cannot consult each person (for a related discussion, see

Finally, we can state:

Loss averse utilitarianism (LAU). *For any* $(z_i^1,...,z_i^m)$ *and any time* t*, if* $SQ_t = (z_j^1,\ldots,z_j^m)$ *then*:

$$V_t(z_i^1,\ldots,z_i^m) = \varnothing\left(\sum_{l=1}^{m} z_i^l - \sum_{l=1}^{m} z_j^l\right)$$

Less formally, the value of any population at time t, according to loss averse utilitarianism, is found by comparing its aggregate wellbeing to the aggregate wellbeing of the 'status quo population' at time t (that is, the actual population at time t), in a way that inflates the difference just in case the status quo population offers higher aggregate wellbeing. This implies that loss averse utilitarianism will, at any time, agree with ordinary utilitarianism—which we can understand as the special case of loss averse utilitarianism when $\varnothing(x) = x$ for any x—on the ranking of all *risk-free* outcomes. But as we shall see in the next subsection, loss averse utilitarianism may disagree with utilitarianism about how to rank *gambles*, even if, say, expected utility theory is assumed as the theory for managing risk.

2.2 First argument: bundling gambles

Let's now turn to the first argument that longtermism supports social risk-taking. The argument is based on the observation that sometimes a sufficiently large bundle of independent gambles of a particular type is acceptable even when an individual gamble of that type is not.

I will continue to focus on 50/50 gambles between someone losing 1 unit of wellbeing and gaining 2 units. For illustrative purposes, I will assume that the value of the status quo (at the time of decision) is 0, both when applying the risk averse (RDU) and the loss averse (LAU) theory. This allows us to determine whether a gamble is worth taking, according to each theory, by checking whether the gamble's value is positive, according to that theory.

Let's start by considering a social planner who applies rank-dependent utility theory to manage risk. Such a planner can evaluate risk-free outcomes various ways, but again I will work with a utilitarian outcome axiology to keep things simple. Given the stipulation that $r(x) = x^2$, a utilitarian who uses rank-dependent utility theory to manage risk will turn down the single gamble, since, as shown above, its value is $-\frac{1}{4}$. Such a decision-maker will

Buchak 2017 and Thoma 2023). And we can understand a social planner as making choices on behalf of the generation that selects them to do so (at least in a democratic society). One way to be cautious when taking risks on behalf of *the generation*, interpreted as a unified agent, is to be more concerned with avoiding losses in *its* wellbeing than with corresponding gains.

also turn down a package of two such gambles, but they will accept a package of three such gambles, since its value is:[17]

$$-3+3*\left(\frac{7}{8}\right)^2+3*\left(\frac{4}{8}\right)^2+3*\left(\frac{1}{8}\right)^2=0.09375$$

Similarly, a loss averse utilitarian who uses expected utility theory to manage risk will, given the above choice of $\varnothing$, turn down the single bet, whose value to them is $-\frac{1}{2}$; they will, however, be indifferent between accepting and rejecting a package of two such bets; and they will accept a package of three such bets, whose value to them is:

$$-\frac{9}{8}+\frac{9}{8}+\frac{6}{8}=\frac{6}{8}$$

To appreciate the importance of the assumption of probabilistic independence—for instance, the assumption that the probability that any gamble turns out unfavourably is independent of how any other gamble turns out—consider first the extreme case where the gambles are all perfectly positively correlated; that is, either they all turn out well or all turn out badly. In that case, the bundle of three gambles is a 50/50 gamble between losing 3 units of wellbeing and gaining 6 units, and a rank-dependent utilitarian and loss averse utilitarian will respectively evaluate the bundle as follows:

$$-3+0.25(9)=-0.75$$
$$0.5(-9)+0.5(6)=-1.5$$

In other words, both the loss averse utilitarian and the utilitarian who uses rank-dependent utility theory to manage risk will turn down the bundle of three gambles when the gambles are perfectly correlated. Now, the same is not true of all instances of *imperfect* positive correlation, that is, cases where some gambles are probabilistically dependent without being perfectly positively correlated. How much the gambles can be positively correlated for the argument to go through depends on details about both the gambles and the degree of risk or loss aversion of the decision-maker. But the general point is that the argument that risk and loss averse decision-makers become more prone to accepting a risky gamble when it is part of a larger bundle of gambles does not hold if the gambles are *too* positively correlated.

What if gambles are *negatively* correlated. In the extreme case, where two gambles of the kind we have been considering are perfectly negatively correlated, the gambles together offer a sure gain of 1 unit of wellbeing, since one gamble will result in a gain of 2 units while the other gamble results in a loss of 1 unit. Therefore, both loss averse and risk averse social planners will accept the bundle of two such gambles. More generally, since negative

[17] In fact, in the limit, with an *infinite* number of gambles, a risk averse maximiser of rank-dependent utility behaves exactly like an expected utility maximiser (see, e.g., Buchak 2013: 217–218).

correlations reduce the possible spread in outcomes, both loss and risk averse decision-makers will typically welcome such correlation.

But returning now to (sufficiently) independent gambles, the upshot of the above is that whether the social planner is risk averse or loss averse, they will want to turn down a gamble like the one we are concerned with here if offered only one such gamble, but they will accept a bundle of three such gambles.[18] And, of course, the same is true of larger bundles, as long as they do not risk a catastrophe or 'extinction' (more on which below).[19]

Now, since the aim of this chapter is to explore what impact a *long-term perspective* should have on social risk-taking—compared to a short-term perspective—the interesting question, however, is not what bundles of gambles a decision-maker should accept when offered at the same time, but rather when offered in sequence over time. Most social planners presumably expect that over their time in office they will be faced with a number of independent gambles that each have a positive objective expectation but which they would nevertheless be tempted to turn down when viewed in isolation. The presumption in favour of longtermism might then seem to imply that such planners should judge each gamble in light of the gambles that they expect they—and perhaps even subsequent social planners—will be faced with in the long run.

However, some subtle philosophical issues now arise, concerning rational commitment and planning. Recall that we are assuming that the planner prefers to turn down each gamble when viewed in isolation. But then given that at each point in time, a decision-maker can arguably at most choose to accept or reject the gamble on offer *at that time*, we can view the decision-maker in question as being faced with a *non-cooperative game* with different *time-slices* of herself. But then the only Nash-equilibrium—that is, the only outcome where no time-slice can do better for itself in light of what others do—is one where each time-slice turns down the gamble with which it is faced. So, we reach an outcome analogous to what Hardin (1968) famously called the 'tragedy of the commons' where all gambles are rejected. (For a more detailed version of this argument, see Stefánsson 2023.)

Here is a different way to arrive at the above conclusion. Suppose that the decision-maker in question predicts that they will know when the *last* gamble on offer arrives. At the start of their time in office, they may predict that they will turn down that last gamble on offer; perhaps because they don't trust their successor to accept similar gambles. Knowing this, they predict that they will turn down the penultimate gamble, and so on, all the way to the first gamble. So, by *backward-induction*, they reason themselves into rejecting all the gambles (cf. Samuelson 1963). Note, however, that this conclusion does not follow if the decision-maker predicts that they will never believe that a gamble on offer is the last on they will face.

It might also be possible to avoid the conclusions of the last two paragraphs by assuming that the decision-maker has fully internalised longtermism. In both paragraphs, the assumption was that the decision-maker sees themselves as having preferences not only about the outcome of the sequence, but also about the outcome at each point in time; moreover, they see themselves not as making one decision for the long-term, but several

[18] As Samuelson (1963) observed, the same is not true, however, of a 'standard' expected utility maximiser, for instance, someone who thinks of monetary outcomes in terms of their own terminal wealth and whose preferences satisfy the axioms of expected utility theory. Such agents will turn down the single gamble, if offered just one, just in case they would also turn down an offer of any finite bundle of gambles.

[19] For a fascinating application of this logic to the question of how to design ethical AI, see Thoma (2022).

decisions that together have a long-term effect. Maybe that is not what it means to internalise longtermism. Perhaps we can instead assume that a longtermist social planner would decide, at the start of their tenure, to accept gambles like those under consideration—due to their prediction about many similar gambles being offered later—and would stick to that decision at each point in time. Now, I am not sure that this assumption is plausibly implied by longtermism. But this assumption (or something like it) is in any case needed for the argument under consideration to establish that longtermism makes a loss or risk averse social planner more risk prone.[20]

If the above is what it means to internalise the longtermist perspective, then a longtermist who is either loss or risk averse (in the sense defined above) is what is called a *resolute chooser* (see, e.g., McClennen 1990).[21] A resolute chooser sometimes resolves to follow a plan and does so even if that means choosing counter-preferentially at some points in time. Some find such counter-preferential choice to be irrational (e.g., Stefánsson 2023). However, it is far from irrational to set up institutions and structures or to pass laws that bind a social planner, in the sense that it removes options that would otherwise be tempting. Nor is it unusual. For instance, the decision by lawmakers in many countries to pass laws that make it illegal for them to meddle with the central bank's interest rate could be seen as an example of a self-binding law, while constitutions can be seen as examples of a social structure that binds lawmakers.

2.3 Second argument: taking other risks into account

Let's now turn to the second argument that longtermism leads to social risk-taking. This argument is based on the observation that in terms of total risk exposure, an individual gamble makes a greater difference when evaluated from the perspective of a risk-free status quo than when evaluated in light of all the risks with which one will be faced.[22] So, in this argument, the importance of taking the long-term perspective does not consist in the fact that the risky *options* with which one will be faced in the future may affect the value of the option with which one is faced today. Rather, this time the long-term perspective is important due to the risks with which one predicts one will inevitably—whatever choices one makes—be faced. Therefore, the subtle philosophical issues about rational commitment and planning that I discussed above do not afflict the second argument.

I will continue to focus on 50/50 gambles between losing 1 unit of wellbeing and gaining 2 units, but now, instead of assuming that the social planner expects to be faced with multiple gambles of that kind, suppose that the social planner evaluates the gamble in light of all the risks with which they predict society will be faced (in whatever time frame they consider relevant). Suppose that the social planner *expects* that the total wellbeing over the time-period with which they are concerned is 100 units. The precise number is of course

[20] This raises the interesting question of whether there is any limit to how many gambles the longtermist should bundle, and, more generally, whether there is rule for deciding how many and which gambles to bundle. I thank David Thorstad for raising this question, which sadly I do not have a good answer to.

[21] See also Thoma (2019) for a useful discussion of resoluteness in the type of decision-problems under consideration.

[22] The argument in this subsection is inspired by Thoma and Weisberg (2017). (They, however, only discuss risk aversion, not loss aversion.)

more or less arbitrary; it just needs to be relatively high compared to the potential loss from the evaluated gamble. Suppose that the planner's expectation is based on a normal distribution around the mean. This assumption of a normal distribution is not essential but simplifies the argument.

For illustrative purposes, suppose that the lowest aggregated wellbeing that the planner considers possible over the relevant time-period is –5 and the highest aggregated wellbeing that the planner considers possible over the relevant time-period is 205. However, the argument would hold even if we increased the numbers in both directions. Perhaps most importantly, the argument would also work if we assumed that the planner thinks that things *could* go *very* much more badly. Finally, to keep the calculations both tractable and illustrative, let's just work with two deviations in each direction from the mean when evaluating the status quo, which corresponds to five deviations in either direction when evaluating the 50/50 gamble. (But the argument can of course be made by assuming a continuous rather than a discrete distribution.) Table 28.1 then represents the status quo and the gamble.

Recall that both the loss averse utilitarian and the utilitarian who uses rank-dependent utility theory to manage risk will turn down a 50/50 gamble between losing 1 unit of wellbeing and gaining 2 units of wellbeing when the gamble is evaluated in isolation. However, as I shall now demonstrate, both types of decision-makers will accept the gamble when evaluated in light of the assumed expectation of a favourable but risky future, as represented by Table 28.1.

Let us this time start by considering loss aversion. How a loss averse utilitarian *now* evaluates this future, with or without the gamble, depends on the *current* sum of wellbeing.

Table 28.1 Distribution of possible outcomes given, on the one hand, the predicted status quo and, on the other hand, the predicted status quo plus accepting the gamble.

		Accept gamble	
Probability	**Status quo**	**Win**	**Lose**
0.025	–5	–3	
0.025	–5		–6
0.1	0	2	
0.1	0		–1
0.25	100	102	
0.25	100		99
0.1	200	202	
0.1	200		199
0.025	205	207	
0.025	205		204

To keep things simple, let us set the current sum at 0; for the purpose of the argument, the only thing that matters is that the current sum is low compared to the most likely future sums of wellbeing. A loss averse utilitarian (as formalised in subsection 2.1) who uses expected utility theory to manage risk will then judge the predicted future without the gamble as follows:

$$205(0.05) + 200(0.2) + 100(0.5) + 0(0.2) - (5 * 3)(0.05) = 99.5$$

But they will evaluate the future with the gamble as follows:

$$207(0.025) + 204(0.025) + 202(0.1) + 199(0.1) + 102(0.25) + 99(0.25)$$
$$+2(0.1) - (1 * 3)(0.1) - (3 * 3)(0.025) - (6 * 3)(0.025) = 99.85$$

So, they will accept the gamble when evaluated in light of their belief that the future will most likely be better than today.

Let's then consider a standard (i.e., not loss averse) utilitarian who uses rank-dependent utility theory to manage risk. They will evaluate the future without the gamble as follows:

$$-5 + 5 * 0.95^2 + 100 * 0.75^2 + 100 * 0.25^2 + 5 * 0.05^2 = 62.025$$

But they will evaluate the future with the gamble as follows:

$$-6 + 3 * 0.975^2 + 2 * 0.95^2 + 3 * 0.85^2 + 97 * 0.75^2 + 3 * 0.5^2$$
$$+97 * 0.25^2 + 3 * 0.15^2 + 2 * 0.05^2 + 3 * 0.025^2 = 62.27375$$

So, the risk averse too will accept the gamble when evaluated in light of their prediction about the risky but, in expectation, favourable future.

In other words, both the loss averse and the risk averse will accept the gamble when evaluated in light of their belief that the future will most likely be better than today but could be worse. And surely that is the belief that a reasonable social planner who takes the long-term perspective would have. We have very good reasons to believe that there will be continued economic growth for the foreseeable future, despite climate change and other potential catastrophes; but we should surely be open to the possibility that things will not go so well.

It may, however, be worth noting that even without allowing for such negative possibilities, a loss averse utilitarian who takes the long-term perspective and who predicts social and economic development to continue will accept the gamble in question when viewed in light of their predicted future (while rejecting it when viewed in isolation). For instance, if they are *sure* that the total wellbeing over the relevant time-period is 100, then they will evaluate the gamble by its expected value. The same is not true of an rank-dependent utilitarian: if they are *sure* that the aggregate wellbeing over the relevant time-period is 100, say, then they will turn down the 50/50 gamble, since then the rank-dependent utility of the future without the gamble is 100, while the rank-dependent utility of the future with the gamble is $99 + 3 * 0.5^2 = 99.75$.

The picture may look different if the social planner expects the future to be *worse* than the present. One complicating factor, however, is that those who are risk averse when they are gambling at or above what they see as their relevant reference point (which often corresponds to their status quo), often become risk *seeking* when they find themselves below that reference point; this is in fact one of the psychological phenomena that Kahneman and Tversky's (1979) prospect theory was designed to capture. So, a social planner who sees their *current* status quo as the relevant reference point, but expects the future to be worse—for instance, in the extreme case, sees human extinction on the relevant horizon—may become risk seeking. And similarly for those who are loss averse when gambling at, above, or below what they take to be the relevant reference points. In contrast, if either a risk or loss averse social planner of the kind we have been considering assumes the future to be *precisely as good* as what they take as the relevant reference point, then they will turn down the single gamble, even when evaluated in light of their prediction about the future, for the same reason as why they turn down the gamble when considered in isolation.

Let me summarise the findings of this whole section before turning (in the next section) to discussing potential policy implications. We have seen that a policy maker who is risk or loss averse, and therefore turns down some gamble that has a positive expectation of the quantity of interest (in this case, wellbeing) if the gamble is viewed in isolation, will not turn down the very same gamble when it is viewed either in light of a favourable (albeit risky) future or in light of sufficiently many additional gambles of the same kind. Moreover, I have proposed that what it means to internalise 'longtermism' is to view each gamble in light of what one expects of the long-term future. But then we seem to have an argument that a longtermist social planner should be more risk prone than a planner who evaluates risks, gains, and losses exactly as the longtermist planner does but nevertheless takes a more short-term perspective.

We have also seen that the arguments in question don't hold for *all* gambles with a positive objective expectation. In the next section, I briefly discuss the conditions needed for the arguments, and compare them to some real-life risky public policies.

3 From theory to practice

Let's begin by considering the probabilistic independence assumption (which was required for the first formal argument) in relation to real-life policy decisions. This assumption will clearly not be true of all 'gambles' with which a social planner is faced. For instance, suppose that a social planner is considering lifting restrictions due to COVID-19 from preschools. We can think of this as a gamble and we can suppose that the social planner sees it as a 50/50 gamble between gaining a benefit of magnitude x and losing a benefit of magnitude $0.5x$; so, the gamble offers an expected benefit. Another similar gamble might be to lift COVID-19 restrictions from high schools. Again, we can suppose that the social planner sees it as a gamble with the same structure as the gamble to lift restrictions from preschools. But the probabilistic independence assumption will presumably not be satisfied with respect to these two gambles. The conditional probability that the gamble to lift restrictions from preschools turns out unfavourably given that the gamble to lift restrictions from high schools turns out unfavourably is presumably greater than the unconditional probability that the gamble to lift restrictions from preschools turns out unfavourably—unless, of

course, we are already certain how the policy change will turn out. So, despite the previous argument, we haven't seen a reason why bundling these and similar gambles should make the social planner more risk prone than if they considered each gamble in isolation.

The probabilistic independence assumption will, however, plausibly be satisfied between many other policies. For instance, suppose that the social planner is considering raising the top tax rate, and let's imagine that they see it as a risky gamble with a positive expectation (say, in terms of aggregate wellbeing). Let's, moreover, suppose that the social planner is considering full legalisation of recreational drugs, and also considers this to be a risky gamble with a positive expectation (in terms of aggregate wellbeing). In this case there will presumably be sufficient probabilistic independence between the two gambles.[23] For instance, the conditional probability that raising the top tax rate turns out unfavourably given that full legalisation turns out badly is presumably (at least roughly) the same as the unconditional probability that raising the top tax rate turns out unfavourably.

Now, it is worth noting that there is a tension between, on the one hand, the probabilistic independence condition and, on the other hand, one of the potential benefits of taking social gambles, or doing 'social experimentation', namely, the information that is gained from both successful and unsuccessful experiments (Barrett and Buchanan 2023).[24] For instance, if (contrary to what was assumed above) the social planner were to *predict* (correctly or incorrectly) that, say, if legalising drugs turns out badly then that teaches them something very important about how people react to incentives, then there might not be sufficient probabilistic independence between the gamble to legalise drugs and the gamble to raise the top tax rate.

Nevertheless, in practice, there certainly seem to be realistic 'social gambles' between which there is sufficient probabilistic independence such that a risk or loss averse social planner would find each gamble to be 'too risky' if considered in isolation, but should nevertheless find the bundle of gambles to be acceptable when considered together. Therefore, if internalising longtermism—or simply taking the long-term perspective—means that the social planner evaluates such gambles together rather than individually, then it follows that there are some real-life examples of risky policies that a longtermist social planner would implement, even though the very same policies would not be implemented by another social planner with the exact same attitudes towards risks, losses, and risk-free outcome, but who takes the short-term perspective (in the sense of evaluating each gamble in isolation).

The example of full legalisation of recreational drugs is useful for exploring some of the other conditions needed for the arguments that longtermism leads to more risk-taking. In the introduction, I gave an informal argument that a social planner might find such an experiment to be unacceptably risky when only considering the impact on the current generation, but would think it acceptable when viewed with a long-term perspective, both because of the potential benefits in the long run if the experiment turns out well, and also because of the knowledge gained even if the experiment turns out badly. Now, although it might be tempting to think of such an experiment as a bundle of gambles, one gamble for each generation, the argument that bundling should make a risk or loss averse decision-maker more prone to accepting the risk clearly doesn't apply to this case, since, for instance,

[23] If the reader disagrees, then consider instead, on the one hand, offering free cancer screening to everyone over 50 and, on the other hand, switching from left- to right-hand driving.

[24] I am grateful to Jacob Barrett for a very useful discussion of this issue.

the probability that the experiment is harmful to one generation is not independent of how the experiment turns out for another generation.

The condition that I, however, want to use this example to illustrate is the qualification that the experiment can be stopped if harmful, and that it is, moreover, predicted that it *will* be stopped if harmful. If that is not the case then it is unclear that the gamble should be considered less risky given a long-term perspective, since then it is a gamble with a potentially huge loss over the long term.

How does this last condition map onto the formal arguments? The reason the condition is satisfied in the formal arguments is that it was assumed that the potential downside from each gamble was rather small in comparison to the total expected outcome (and also in comparison to the most likely outcomes); in the first formal argument the total expected outcome is from the bundle of gambles, whereas in the second formal argument the total expected outcome is from the predicted future (with or without the gamble). There are, however, examples of real-life public policy choices where this condition would not seem to be satisfied.

Let's take a historical example. In the 1990s Sweden saw a radical social experiment where the education sector was completely opened to private actors: while the state would continue to fund education, anyone who fulfilled some minimal criteria could open a school and provide state-funded education. At the same time, parents were given (in principle) complete freedom to choose schools. The hope, of course, was that competition between schools—both private and public—combined with parents' preference for getting the best education for their children, would result in a situation where only those education providers who could offer the best education for the budget given by the state would survive. (A similar experiment was made in various other areas, for instance, in the health-care sector.)

Today most experts seem to agree that overall the experiment has been harmful. The average quality of primary and secondary education seems to have diminished—as judged by students' performances in international tests—while the spread in education outcomes has increased (Molander 2017). Moreover, the public seem to be quite unhappy with the current system; for instance, a majority of Swedes are in favour of a cap on the level of dividends private education providers can pay their owners (see, again, Molander 2017). But despite this popular and expert opinion, the proportion of private providers has only increased (Molander 2017) and political attempts at capping dividends have so far proven mostly unsuccessful. One perhaps obvious reason for this is all the lobbying done on behalf of private education providers (see, e.g., Svallfors and Tylström 2017). The experiment created a financially very strong interest group that of course does whatever it can to maintain the current status quo.[25]

Coming back to the formal arguments from the last section, the observations from the last paragraph may show that the experiment to open up the education sector to private providers should not be seen as a gamble where the potential downside is not really significant in comparison to the expectation about the long-term. If the harmful effects that seem to have resulted from this experiment get reproduced for each subsequent

[25] I speculate that another reason why it has proven hard to discontinue this radical experiment is that it is generally much harder to implement policies that are seen as removing a freedom to choose than it is to implement policies that are seen as increasing the freedom to choose.

generation of school children for a very long time, then the actual long-term downside is in fact very significant.

A general lesson that emerges from the above social experiment is that the argument from longtermism to social risk-taking may not work in cases where a gamble is likely to create influential groups who have an interest in maintaining the resulting status quo irrespective of the social consequences. This would seem relevant when reasoning about, for instance, experimenting with legalising recreational drugs. Such legalisation would presumably prove quite profitable for some legal companies (e.g., today's pharmaceutical companies) that would then have an interest in maintaining the resulting status quo—unless, of course, private actors would not be allowed to profit from the production or sale of recreational drugs. More generally, when assessing whether taking the long-term perspective supports accepting a particular social gamble, one must ask whether the potential downside of the gamble is likely to be reproduced for the coming generations.

Finally, and to conclude, I will say a few words about the condition that the gambles in question have no chance of causing a *catastrophic* outcome (in particular, extinction). For instance, in the first formal argument, about bundling gambles, if losing some gamble(s) in the bundle means that the decision-maker will not be in a position to accept further gambles, then the argument in question would not be applicable. The same is true if losing one gamble means that one won't enjoy the fruits of winning a future gamble. Similarly, in the second formal argument, about the expectation of a favourable future, if the gamble under evaluation is sufficiently large-scale such that losing it means that the future is unlikely to be favourable, then that argument for risk-taking does not apply. This means that the arguments in question do not apply to gambles with potentially catastrophic outcomes. In other words, the formal arguments that show that, under some conditions, longtermism supports risk-taking do not show that a longtermist should be less averse to catastrophic risk than a short-termist.

In fact, under some conditions, being a longtermist should make one *more* averse to catastrophic—in particular, extinction—risks (a point made, e.g., by Greaves and MacAskill 2021). As noted above, how a risk or loss averse social planner behaves if they expect human extinction whatever they do, depends on details about how gambling below the relevant reference point affects one's attitudes to risks and losses. But it is straightforward to demonstrate that a risk or loss averse social planner who predicts the future will be favourable, *unless their society goes extinct*, will be less prone to accept gambles that risk extinction than their short-termist counterpart. For a simple demonstration, let's consider a social planner who predicts they will be offered a sequence of the type of 50/50 gambles that have so far been the focus of this chapter. Further, suppose that they in addition currently face a 50/50 gamble between, on the one hand, gaining 4 units of wellbeing and, on the other hand, 'extinction', interpreted as them not being in the position to accept further gambles. Finally, suppose that they don't expect any potential for gains or losses except for, first, those from the aforementioned sequence and, second, from the gamble currently on offer. Whether risk or loss averse (in the sense defined above), the social planner may then accept the extinction-risk gamble if they are so short-termist that their planning horizon only includes, say, the next two gambles in the sequence. For then they may reason that what is lost by extinction is no too great, so they might as well take the gamble. In contrast, if their planning horizon includes the next hundred gambles, say, then they will reject the

extinction-risk gamble, since it threatens preventing them from accepting over time an incredibly favourable bundle of gambles. So, a longtermist who expects the future to be favourable will be more averse to extinction risk than their short-termist counterpart.[26]

References

Barrett, J. and Buchanan, A. (2023), 'Social Experimentation in an Unjust World', in D. W. Sobel and S. Wall (eds.), *Oxford Studies in Political Philosophy Volume 9* (Oxford University Press), 127–152.
Buchak, L. (2013), *Risk and rationality* (Oxford University Press).
Buchak, L. (2017), 'Taking Risks Behind the Veil of Ignorance', in *Ethics* 127/3: 610–644.
Greaves, H. and MacAskill, W.(2021), 'The Case for Strong Longtermism', GPI Working paper 5-2021 (Global Priorities Institute, Oxford University).
Hardin, G. (1968), 'The Tragedy of the Commons', in *Science*, 162/3859: 1243–1248.
Kahneman, D. (2012). *Thinking, fast and slow*. London: Penguin Books.
Kahneman, D. and Tversky, A. (1979), 'Prospect Theory: An Analysis of Decision under Risk', in *Econometrica* 47/2: 263–291.
McClennen, E. F. (1990), *Rationality and Dynamic Choice: Foundational Explorations* (Cambridge University Press).
Molander, P. (2017), *Dags för Omprövning* (*Time for Re-evaluation*) (Expertgruppen för Studier i Offentlig Ekonomi 2017:1) (Regeringskansliet).
Rabin, M. (2000), 'Diminishing Marginal Utility of Wealth Cannot Explain Risk Aversion', in D. Kahneman and A. Tversky (eds.), *Choices, Values and Frames* (Cambridge University Press), 202–208.
Rabin, M. and Thaler, R. H. (2001), 'Anomalies: Risk Aversion', in *The Journal of Economic Perspectives* 15/1: 219–232.
Rothschild, M. and Stiglitz, J. (1970), 'Increasing Risk: I. A Definition', in *Journal of Economic Theory* 2/3: 225–243.
Samuelson, P. (1963), 'Risk and Uncertainty: A Fallacy of Large Numbers', in *Scientia* 98: 108–113.
Stefánsson, H. Orri (2023), "The Tragedy of the Risk Averse", in *Erkenntnis* 88/1: 351–364.
Stefánsson, H. O. and Bradley, R. (2015), 'How Valuable Are Chances?', in *Philosophy of Science* 82/4: 602–625.
Stefánsson, H. O. and Bradley, R. (2019), 'What Is Risk Aversion?', in *British Journal for the Philosophy of Science* 70/1: 77–102.
Svallfors, S. and Tyllström, A. (2017), 'Lobbying for Profits: Private Companies and the Privatization of the Welfare State in Sweden', Institute for Futures Studies Working Paper 2017:1.
Thoma, J. (2019), 'Risk Aversion and the Long Run', in *Ethics* 129/2: 230–253.
Thoma, J. (2022), 'Risk Imposition by Artificial Agents: The Moral Proxy Problem', in S. Vöneky, P. Kellmeyer, O. Müller, and W. Burgard (eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (Cambridge University Press), 50–66.
Thoma, J. (2023), 'Taking Risks on Behalf of Others', in *Philosophy Compass* 18/3: 1–13.
Thoma, J. and Weisberg, J. (2017), 'Risk Writ Large', in *Philosophical Studies* 174/9: 2369–2384.

[26] This chapter has been presented at the 9th Oxford Workshop on Global Priorities Research and at the PPE seminar at the Institute for Future Studies in Stockholm (both in March 2022). I am very grateful for the comments and suggestions I received at these two events. Johanna Thoma read a draft of the chapter and provided very helpful written feedback. Finally, thanks to the editors of this volume and a referee, for useful comments and suggestions.

29

The Short-Termism of 'Hard' Economics

Ilan Noy and Shakked Noy

1 Introduction

The academic economics profession has contributed, and seems poised to contribute, very little to longtermist research. This is despite the following facts: Effective altruist organizations have long recommended an economics PhD as one of the best ways to acquire skills for global priorities research, including longtermist research; many foundational concepts and tools in economics lend themselves naturally to longtermist research, including ideas about constrained optimization, strategic interactions, and intertemporal decision-making; and many individual economists are enthusiastically interested in longtermism.

In this chapter, we attempt to explain this surprising juxtaposition. Our main claim is that the academic economics profession is stuck in a methodological straitjacket that prevents it from accepting more methodologically pluralistic kinds of research—including, as a side effect, potential longtermist research. Our criticism of the profession's methodological narrow-mindedness is not new: we echo George Akerlof's (2020) indictment of 'hardness bias' in economics and point to survey evidence showing that a majority of economists dislike the profession's current methodological norms (Andre and Falk 2021). Our aim with this chapter is threefold. First, for readers of this volume, who are by now familiar with longtermism but may be unfamiliar with academic economics, we give an overview of the state of the profession. Second, we aim to raise the salience and importance of the profession's narrow-mindedness by illustrating how it prevents economists from contributing to longtermist research. Third, we try to articulate a more positive vision for how economists could contribute to longtermist research if current methodological norms were relaxed.

Throughout, we take 'longtermism' to mean the view that the impacts of our actions on the very long-term future deserve prominent consideration in decision-making. Areas of interest to longtermists that we mention in this chapter include catastrophic or existential risks from artificial intelligence, climate change, or nuclear war; long-run rates of scientific innovation and economic growth; and global governance institutions.

The rest of this chapter proceeds as follows. In section 2, we describe the methodological constraints and norms that shape the way economists currently do research. In section 3, we consider *why* these methodological norms have arisen. In section 4, we use three case studies to illustrate how these norms prevent economists from producing research relevant to longtermism. In section 5, we argue that academic economists could contribute substantially to longtermist research if given the opportunity. In section 6, we consider alternative explanations for economists' failure to engage with longtermist

Ilan Noy and Shakked Noy, *The Short-Termism of 'Hard' Economics* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press. © Ilan Noy and Shakked Noy 2025.
DOI: 10.1093/9780191979972.003.0029

research. Finally, in section 7, we offer some speculative recommendations to institutions focused on longtermist research.

2 What does academic economics research look like?

2.1 Norms of typology, exclusion, and omission

The production of academic economics research is tightly disciplined by a set of norms, which we can divide into norms of typology, exclusion, and omission.

Norms of *typology* establish a small set of 'kinds' of economic research and insist that any new piece of research must fall neatly into one kind. Each kind is associated with its own methodology and its own standards for what counts as a valuable (and hence publishable) contribution to economic knowledge. Norms of typology, by forcing new research to adhere rigidly to one predetermined kind, discourage projects that draw on a more varied set of methodologies and combine them in different ways, or venture into entirely new methodological terrains. As we will later suggest, longtermist projects disproportionately share these latter characteristics.

Norms of *exclusion* say that certain types of evidence or research practices are not scientifically acceptable. Often, these are types of evidence or research practices that are widely used in other fields of social science (e.g. focus groups or Delphic surveys). By precluding research that draws heavily on these types of evidence or research practices, norms of exclusion narrow the field of acceptable economics research. Again, this makes longtermist economic research—which is often methodologically pluralistic and draws on insights from multiple fields of social science—difficult.

Norms of *omission* say that certain types of evidence or research practices, while acceptable, are not significantly professionally rewarded. Since economists, like any other researchers, face time and resource constraints, and since career promotion paths force them to concentrate on professionally rewarded activities, norms of omission inevitably cause certain types of research to fall by the wayside—including, as we will describe, policy-relevant research, interdisciplinary research, book-length research, and highly speculative research.

In the following subsections, we describe the key norms of typology, exclusion, and omission that shape the current landscape of academic economics research. We then turn, in section 3, to the underlying fundamental values of the economics profession that give rise to these norms.

Two important caveats apply to our discussion in this section. First, we are not claiming that these norms are wholly unjustified or even harmful on average. As we will see in this section and section 3, there are many good reasons behind these norms, and these norms have many positive effects. Rather, our argument is that the profession currently adheres to these norms too rigidly, and that relaxation of these norms on the margins could beneficially broaden the scope of economics research, including by increasing the profession's receptiveness to longtermist research. Second, we are making broad generalizations about the economics profession; obviously these generalizations are not *universally* true, and there are exceptions to each claim we are making. Nevertheless, we believe that identifying

these very common features is still important in clarifying the potential role of economics in longtermist research.

2.1.1 Norms of typology

Almost all academic economics research can be classified into one of the following three categories: theory, empirical causal inference, and structural modelling and estimation.

A typical *economic theory* paper does one of two things. One approach is to introduce a novel economic mechanism, describe the conditions under which it arises, and illustrate its consequences. For example, the mechanism might be that, when sellers have more information about the quality of a product than potential buyers, a recursive series of buyer inferences about expected product quality and seller decisions can cause a market to 'unravel' and prevent mutually beneficial trades from occurring (Akerlof 1970). Alternatively, an economic theory paper might introduce a conceptual framework for thinking about a particular problem or context and show what can be proven about that problem or context. For example, a paper might provide a framework for thinking about optimal savings rates across generations (Ramsey 1928).

Economic theory papers share two key features. First, there is a focus on precision through mathematical formalization, which avoids the ambiguity that often clouds ordinary language but also results in heavy abstraction. Second, there is a focus on generality: identifying the broadest possible conditions under which a mechanism arises, or keeping a conceptual framework as expansive as possible. Naturally, a focus on generality encourages abstracting away from contextual particulars, even when there is a relatively limited set of contexts for which this focus is meaningful. It also encourages a search for mechanisms or problems that apply very generally (even if they are not that important), and discourages work on mechanisms or problems that may be very important but operate only in narrow particular contexts (e.g. in the context of the European Union's scientific funding model, or China's regulation of artificial intelligence research). This discourages theoretical economists from modelling practical situations or applying general results to concrete contexts.

Empirical economics research generally attempts to estimate the causal effect of a particular shift, change, or intervention (a policy, an event, or a circumstance) on a given outcome. For example, a paper might try to estimate the effects of class size on educational attainment and future earnings, the effects of a neighbourhood's characteristics on the well-being of its residents, or the effects of fiscal and monetary policy tools on inflation. Clear causal inference is the gold standard of most empirical research, and it is frequently the strength of this causal identification (rather than the importance of the question being studied or its policy implications) that is the main yardstick by which journal publication decisions are being made.

A focus on causal effects sounds innocuous, but it narrows remarkably the scope of acceptable (i.e. publishable) empirical research. First, it results in bias against *descriptive*, *explanatory*, or *predictive* empirical research. It is difficult to publish *descriptive* research illustrating an important pattern or trend, or providing suggestive evidence for a causal mechanism that is hard to identify more credibly (think, for example, about work on 'deaths of despair'). Economists almost never formulate general *explanations* of what drives variation in the outcome they study (Rodrik, 2021). For any outcome *Y*, there will be hundreds

of papers asking whether X_i has a causal effect on Y for different choices of i, but very few papers ask which variables explain most of the variations in Y. Finally, generally speaking, economists are not interested in attempts to use empirical evidence to *forecast* important future trends (contrary to the lay perception that economists spend all day trying to predict recessions). All of this is bad news for longtermist research, since it precludes study of questions like 'how has aggregate global welfare changed over time?' 'what are the main drivers of scientific innovation?' or 'what are the most plausible trajectories for economic growth over the rest of this century?'

Moreover, the constraints of causal inference restrict the questions, contexts, and time periods that can be studied. Clean identification of causality is much more likely for narrow policy interventions that can be neatly isolated. This prevents study of larger, more important interventions that are harder to disentangle from other factors and from each other. Economists only rarely try to answer big questions like whether democracy contributes to economic growth (Acemoglu et al. 2019). Moreover, empirical causal inference usually requires very large existing datasets (especially since most research designs require zooming in on a small subset of the data). Most large datasets began their coverage only in the 1980s or 1990s, mostly in high-income countries. Besides, the assembly of new datasets is not adequately rewarded professionally.

Structural modelling and estimation, the third type of paper, occupies a kind of middle ground between the first two types. A typical 'structural' paper sets up a model to describe a particular economic system or context, usually drawing on the economic theory literature on that subject. The model will feature various 'parameters' that could have a variety of values—for example, the price-demand elasticity of a certain class of products, or the discount rate. Typically, the paper will 'plug in' estimates of these parameters from the existing empirical literature on the subject, or produce its own estimates based on some set of assumptions. The model will then be used for any of the following purposes: connecting estimates of behavioural changes to fundamental parameters like time preferences; making statements about counterfactuals (what would have happened if we changed this policy parameter by X); or making statements about welfare (how does this policy affect aggregate welfare?).

Of the three kinds of research, structural modelling is the most methodologically pluralistic and perhaps the closest to what ideal longtermist research could resemble. Papers of this type combine theoretical modelling with estimates from specific contexts; standards for the empirical component of these methodologies tend to be more permissive than those of pure causal inference papers, and the theory is also typically less abstract, sometimes incorporating institutional details. That said, there are still important limitations to the structural modelling approach. Even in very complex modelling (for example, in agent-based models—e.g. Lamperti et al. 2018) the approach requires many structural assumptions about functional forms and parameters, because of the formalistic methodology, and these make those models of questionable use in identifying solutions to real-world problems.

In recent years, there has been a growing expectation that papers should include both theoretical and empirical components. A paper is then viewed either as a primarily theoretical paper with supporting empirical evidence, or a primarily empirical paper grounded in a mathematized conceptual framework. While this trend, an attempt to unify these disparate methodological approaches, has many advantages, it has not fundamentally transformed

the character of economics research. The scope of publishable 'primarily empirical' research remains constrained by the norms we outlined above, and the requirements imposed on 'primarily theoretical' papers also remain the same as above.

2.1.2 Norms of exclusion and omission

Implicit in this typology of papers is that specific kinds of empirical evidence are excluded from consideration in mainstream economics. First, qualitative research as conducted by sociologists or anthropologists—e.g. interviews, case studies, focus groups—is not considered an acceptable source of evidence. Second, non-statistical empirical work of the kind that historians conduct is also dismissed. Pieces of qualitative and non-statistical evidence can be (and often are) invoked by economists to *motivate* theoretical models or statistical research or to supplement discussion of possible mechanisms, but these kinds of evidence can never be the centrepiece of a published research project. This is regrettable even if we agree that quantitative causal inference is superior to other kinds of empirical evidence, because there are *many* questions or contexts that cannot be studied with quantitative causal inference. Although they cannot be studied because of data limitations or lack of easily justified exogenous variation, they may be amenable to alternative forms of empirical investigation. These, again, may include questions of interest to longtermists.

Economics is the least interdisciplinary of the social sciences, by a wide margin. Compared to sociologists, political scientists, and anthropologists, economists are much less likely to cite papers from other disciplines, as they view interdisciplinarity as less important (Fourcade, Ollion, and Algan 2015). Our anecdotal impression is also that economists are much less likely to coauthor with scholars from other disciplines and to publish papers in other disciplines' journals. Since, as we will argue further on, promising longtermist projects often cut across disciplinary boundaries, the insularity of economists and their hostility to interdisciplinary work is a barrier to longtermist research.

In addition, or maybe as a reflection of this disciplinary insularity, economics research published in leading venues rarely cites research written in any language other than English. This is true even for research papers that focus on a non-English-language country.

The standard in economics is to adopt a conservative and buck-passing approach to normative analysis. Economists are usually unwilling to impose substantive and explicit normative assumptions, and typically restrict themselves to 'recovering' normative valuations from the behaviours of market participants. For example, the social discount rate, a parameter central to any longtermist discussion, is calculated from the time preferences implicit in the relative returns of financial instruments; another example is the practice in standard welfare analysis in public economics of using people's willingness to pay for an intervention as a measure of that intervention's benefits (Greaves, 2017; Finkelstein and Hendren, 2020). These approaches are *conservative* in the sense that they take the prevailing and observable normative values (such as people's time preferences) as given and do not question them, and they are *buck-passing* in the sense that they recognize the need to make normative judgements, but are reluctant to do so, arguing that these should be made by non-economist others. Obviously, this approach to normative analysis makes economics reflexively resistant to longtermist projects that are underpinned by unconventional normative assumptions or attempt to engage in welfare analysis without any access to observable data (like willingness-to-pay surveys) that can help pin down the prevailing normative view of these issues.

Another peculiar aspect of academic economics, which leads to equally perverse exclusionary practices, is the dramatically disproportionate career payoffs associated with publishing papers in the 'top 5' economics journals (the *American Economic Review*, *Quarterly Journal of Economics*, *Econometrica*, *Journal of Political Economy*, and *Review of Economic Studies*). Successfully publishing in the top-5 journals increases an economist's probability of getting tenure, prospects for promotion, and status in the profession by an amount that is significantly greater than any other kind of professional achievement (Heckman and Moktan 2020). This results in a strong norm of omission: any work not ultimately oriented towards a publication in the top-5 has much smaller professional payoff. Consequently, a laundry list of activities are implicitly, and often explicitly, discouraged. These include publishing in other disciplines' top journals (e.g. *American Political Science Review*), or even top interdisciplinary outlets such as *Nature* or *Science*. These practices end up producing a high volume of important knowledge that aligns perfectly with the prevailing top-5 norms, but is nevertheless not pathbreaking. Moreover, these top-5 journals are mostly controlled by editors from very few US academic departments, resulting in extreme centralization of power and hence of decisions about which kinds of research are interesting or valuable. This narrows the scope of economics research even further.

The fact that economists rarely write books deserves particular emphasis. Unlike other social sciences, where book-writing is encouraged and often a *de facto* requirement for tenure or promotion, academic economists very rarely write books and produce most of their intellectual output in the form of refereed journal articles. This further contributes to the aforementioned culture of focusing on the identification of particular causal effects or explication of particular mechanisms rather than the production of broad general explanations of important phenomena. A causal effect can be estimated in one paper, but a persuasive explanation or overview of a new and complex problem often demands a book-length exposition.

3 Why does academic economics research look like that?

The norms we just outlined are relatively recent phenomena, having arisen in the past 30–50 years. (Economists prior to this period were much more methodologically pluralistic, interdisciplinary, interested in general explanations, and so on.) Why has the academic economics profession acquired all the norms we outlined above?

George Akerlof's (2020) diagnosis, which we find persuasive, is that the underlying cause of all the aforementioned norms is economists' increasingly strong preference for methodological *hardness* in research. A glib way of defining 'hardness' would be to say that modern economists like to see themselves as sitting near the *hard* end of a spectrum ranging from 'hard sciences' like physics to 'soft sciences' like sociology. More descriptively, a preference for 'hardness' means a preference for precise mathematical formalization over verbal exposition; for 'hard, objective' numerical quantitative evidence over 'soft, subjective' qualitative evidence; for causal over correlational/descriptive quantitative evidence; and so on.

Economists' preference for hardness provides a unified explanation of everything we described in section 2. Papers are divided into three discrete types because each type exemplifies the 'hardest' possible realization of theory or applied work. Economists are insular and disdain interdisciplinary engagement because other social scientific disciplines are less

'hard' so that engaging with them would pollute the quality of economics research. Certain types of evidence are excluded because they are 'soft'. The same goes for substantive normative assumptions, which are avoided because they are perceived as allowing subjective assumptions to slip in. Economists focus on narrow identification of causal effects rather than general explanation of phenomena because the former is more amenable to 'hard' methods. Even the disproportionate professional rewards associated with the top-5 journals arguably reflect a preference for hardness; there is a perception that these journals are uniquely demanding in the rigour they expect.

Of course, there is much to like about a focus on hardness. For example, economists tend to have higher standards than other social scientists for drawing inferences about causal effects from statistical evidence, and are much less likely to over-infer causality from correlational evidence. Economists are also relentlessly precise about isolating and stating theoretical mechanisms in a way that other social sciences are not.

But even initially beneficial preferences can go too far, and we agree with Akerlof that the economics profession currently places an *excessive* premium on hardness. Many economists concur; Andre and Falk (2021) report that a majority of economists would prefer to see more interdisciplinary research, more applied research, more risky and disruptive research, and more research on important questions even at the expense of the quality of causal identification—all of which entails a partial sacrifice of hardness.

What, exactly, is the problem with too much hardness? As Akerlof describes it, the problem is that an excessive focus on hardness means the economics profession fails to balance the 'hardness' and 'importance' of potential research projects. A balanced research portfolio would include projects that are amenable to hard methods, meaning ones that can produce very precise and confident answers to those questions, even if the questions are ultimately not very important; *and* it would include projects on important questions, even if those questions are not very amenable to hard methods, so that we must make do with the evidence that can feasibly be mustered. Unduly strict demands for hardness *rule out* the latter kind of projects and mean that the economics profession neglects many important questions simply because they are unsuitable for study using the 'hard' methods the profession prizes above all else.

Obviously, *many* longtermist projects fall straight into the 'important but difficult to study with hard methods' bucket. Questions about the long-run drivers of scientific innovation, or ways of structuring global governance to prevent catastrophic risks, are incredibly important but difficult to study with hard methods (governance of Artificial Intelligence being a prime example of such a topic). They are difficult to study with hard *theoretical* methods because making progress on these questions presumably involves analysis of very concrete particular facts about existing or historical institutions and events, rather than the exposition of new abstract and general mechanisms. And they are difficult to study with hard *empirical* methods: they ask prospective rather than retrospective questions, and require answers to a combination of descriptive, explanatory, predictive, and causal sub-questions.

There is another way of stating the problem—one that ties neatly into the *epistemic* culture of longtermism. The hardness paradigm imposes a kind of dichotomy onto pieces of empirical evidence: either this piece of evidence meets the commonly accepted standards for causal inference (i.e. having a recognizable identification strategy that depends on plausible identifying assumptions), or it does not. In the latter case, it is thrown out completely

and disregarded (or published in 'inferior' journals). But this approach is incompatible with the subjective Bayesian approach popular among longtermists, which acknowledges that *any* kind of observation can be informative about a hypothesis. In particular, subjective Bayesians recognize that for many very important questions, (i) we have very little relevant evidence to begin with, (ii) credibly identified causal evidence is probably unattainable for now, and (iii) as a consequence, new pieces of correlational, descriptive, or qualitative evidence can be *very* informative and valuable. Many, if not most, economic questions relevant to longtermism fall into this category; yet because they cannot be studied using standard methodologies and get published in economics journals, they are rarely pursued.

This is not to say that *no* research relevant to longtermism can fit within the confines of the economics profession's existing norms. In the next section, we will consider what is possible within the current confines of economic research, and what is not. We do so by considering three fields of inquiry closely related to longtermist considerations: frameworks for thinking about long-term decision-making; climate change; and artificial intelligence.

4 Academic economics and longtermist research topics

We now survey some areas of research relevant to longtermism and discuss economists' limited contribution to them. The first topic—conceptual frameworks for thinking about the long-term—is intrinsically amenable to hard methods and therefore economists have contributed significantly to it. By contrast, for the second and third topics—climate change and artificial intelligence—much less has been done and what has been done is narrowly focused on topics that are amenable to the methodological preference for hardness. The two topics are different. Climate change, as a focus of research, has generated voluminous work in many disciplinary fields for several decades, while artificial intelligence is a much newer topic in general, and for the social sciences in particular. Still, in both cases, the contribution of economics has been quite limited.

4.1 Conceptual frameworks

At least since the pioneering work of Frank Ramsey in the 1920s, economists have been at the forefront of producing rigorous frameworks that permit analysis of questions relating to long-term resource allocation, welfare analysis, and existential risk. These are areas where intuitions and qualitative arguments break down very quickly, and where economic theory really shines by providing elegant mathematical models with simple foundations that flesh out, systematize, and extrapolate from basic intuitions and assumptions. Economic theorists have analysed questions about optimal intergenerational savings (Ramsey 1928), evaluation of social welfare across time and generations (Dasgupta 2019), the long-run drivers of economic growth, and the importance of catastrophic and existential risks (Millner and Heal 2023), among others.

Often, despite making minimal assumptions, these frameworks produce substantive recommendations, in addition to helping us think more clearly about specific questions. For example, they might tell us that catastrophic risks (with non-trivial probabilities) make standard cost-benefit analysis break down (Weitzman 2009; Millner 2013), that the

primary levers to affect long-run economic growth are related to the rate of innovation (Romer 1990), or that economic growth cannot be sustained in the long run with a declining population (Jones 2022).

In this area, we think academic economics has done and continues to do well. The tools of economic theory are extremely well suited for tackling these sorts of questions, and the profession appropriately rewards the development of elegant theories illustrating important abstract mechanisms and hypotheses. As long as economists are aware of the importance of questions about humanity's long-term future—which we think they increasingly are—there should be few barriers to conducting this kind of conceptual longtermist research.

However, these conceptual frameworks and simple recommendations only get us so far. They may tell us which parameters likely govern the optimal savings rate or identify the welfare trade-offs between the present and the distant future, but they do not let us straightforwardly estimate those parameters. They tell us that the innovation rate is important but offer little guidance on how to increase it. They impress on us the importance of reducing catastrophic risks but offer little guidance on the sources of risk, or on how to reduce risks. When we come to those more practical policy questions, hardness bias causes economics to do even less well. Climate change and artificial intelligence belong exactly to this group of policy-relevant research topics.

4.2 Climate change

Climate change research is of interest to longtermists for a number of reasons. First, while there is a majority view among longtermists that climate change is unlikely to directly cause human extinction, there is still substantial uncertainty surrounding that assessment. Second, climate change might be a contributing risk factor for other genuine sources of existential risk such as a nuclear war between the Great Powers. Third, climate change research is the largest and most mature research field dedicated to a forward-looking, long-term topic; it therefore offers a convenient window into what longtermist research that is not focused exclusively on existential risks might look like.

Economists are naturally not expected to contribute to research on the physical science of climate change. But two areas of climate research cry out for economic expertise: research on the economic and socio-economic consequences of anthropogenic climatic changes, and research on ways to reduce greenhouse gas emissions or adapt to the consequences of any residual climatic changes. These areas of research correspond roughly to the Intergovernmental Panel on Climate Change's (IPCC) Working Group II, on vulnerability, impacts, and adaptation, and Working Group III, on mitigation (henceforth, WG-II and WG-III).

Below, we argue that the hardness norms we identified above are responsible for economists' lack of engagement with the issues arising because of anthropogenic climate change. These norms have led economists to make narrow contributions that have alienated researchers from other disciplines and led to the side-lining of the profession from the literature (Noy and Uher, 2022; Noy, 2023). Moreover, we argue that hardness biases have led economists to neglect considerations that are of special importance to longtermists, specifically extreme tail risks from climate change (such as those that can arise from tipping points and cascading impacts).

4.2.1 Working Group II

The official mandate of WG-II is to 'assess the vulnerability of socio-economic and natural systems to climate change, negative and positive consequences of climate change and options for adapting to it'. The engagement of economists with WG-II has mostly focused on Integrated Assessment Models (IAMs), the workhorses of climate change economics since the pioneering work of William Nordhaus (1975; 1992) and Nicholas Stern (2007). These models have been widely criticized for underestimating the risks of climate change (Howard and Sterner 2017; Stern and Stiglitz 2022). This underestimation is a direct consequence of hardness bias. In Nordhaus's pioneering IAM (the Dynamic Integrated Climate-Economy [DICE] model), the link between climate and the economy is modelled by the equation: $D = \alpha\Delta T^2$—with D defined as the damage sustained by the global economy from climate change and ΔT denoting the difference between global average temperature today and in preindustrial times. The parameter α is an aggregator that summarizes the various channels through which climate change has an impact on economic activity (it is mysteriously parameterized as $\alpha = 0.00236$). This setup incorporates no uncertainty around the magnitude of α and does not allow for extreme tail risks, resulting in 'profoundly misleading' underassessment of climatic risks (Stern 2013, p. 839).

The α shortcut exemplifies the methodological straitjacket economists operate under. The shortcut is motivated by the demand for mathematical tractability and simplicity. The lack of attention paid to the assumed magnitude of α is a consequence of economists' lack of engagement with other disciplines (like the physical sciences, or the sociology of disasters), as well as their reluctance to engage with messy empirical questions that cannot be tackled with 'neat' causal methods. Meanwhile, the failure to accommodate tail risks stems from the constraints of structural modelling and estimation. IAMs cannot, as a rule, accommodate extreme tail risks, nor can they estimate the key parameters involved in modelling phenomena that have not happened repeatedly and whose likelihood cannot be estimated from retrospective data. Economists' refusal to engage with subjective data (like aggregations of expert forecasts of these risks) means that these risks and phenomena are left out altogether.

Of course, economic questions relating to climate change are very difficult to answer. It is challenging enough to model climatic systems featuring non-trivial likelihood of abrupt changes and tipping points; additionally modelling society–climate interactions, with their own tipping points, irreversibilities, and multiple equilibria, is doubly complex. But the self-imposed constraints of economists' research practices mean they do not seriously engage with the complexity and contextual particulars of these problems and appear to prefer abstractions that render their conclusions unconvincing to those working outside of economists' disciplinary boundaries.

The other approach adopted by economists to examine the impacts of climate change is based on empirical causal inference from, typically, country-level macroeconomic data, collected across both different geographies and different times. These types of investigations can provide identification of the causal impact of the climate by, basically, assuming that weather measurements are exogenous to the economic system (e.g. Dell, Jones, and Olken 2012; Hsiang et al. 2017; Kalkuhl and Wenz 2020). This same approach, of 'looking back to see better ahead', is also used in the few papers that have looked at climate change adaptation practices (e.g. Burke and Emerick 2016; Chen and Gong 2021).

These papers use economists' standard methods of empirical causal inference. Again, the inherent limitations of these methods mean that these papers neglect important

considerations that have little trace in recent historical data, including catastrophic tail risks. These backward-looking papers also rely on reduced-form partial-equilibrium approaches, again limiting their scope. Specifically, this literature is largely unable to consider a crucial scenario—the low likelihood that current impacts will make a major difference to very long-term economic dynamics. This typological straitjacket has thus largely prevented economists from working on extreme risks, but it also prevents them from asking other questions, such as the impact of social cascades following weather extremes (such as was hypothesized for the recent Syrian Civil War (Ide 2018)).

4.2.2 Working Group III

Economic research is equally constrained on the topic of mitigation (WG-III), as it mostly focuses on two issues: quantifying the social cost of carbon, and questions around the (mechanism) design of emission-permit markets and carbon taxes on various greenhouse gas emissions. The calculation of the social cost of carbon, a central quantity, requires an assessment of the impact of climate change (the remit of WG-II), so the description of constraints described above applies here as well. Within the context of WG-III's discussion of mitigation, the widely perceived underestimates of the social cost of carbon have angered researchers from other disciplines, and have recently led to the marginalization of economics from the mitigation literature within the IPCC (Chan et al. 2016).

Reliance on formalized economic theory has led most economists to view the emissions problem as a classic negative externality (i.e. when a consequence of an action is not part of the consideration set of the actor). Hence the near-unanimous conclusion has been that the preferred policy tool to prevent and mitigate climate change should be a carbon tax (Timilsina 2022). Economists have been especially blind to the near-total political infeasibility of adequately high carbon taxes. In a paper summarizing economists' views on climate policy, Hassler et al. (2016) claim, in spite of the overwhelming evidence to the contrary, that 'at this moment in time, we judge a carbon tax to be politically feasible. One often hears that carbon taxes are politically infeasible; we argue that they are likely not' (p. 506).

It would be unfair to give an entirely negative view of economists' work on these topics. A small number of researchers do, for example, engage better with the issues of catastrophic risks—issues that relate to our discussion in the previous section on conceptual frameworks (economists typically model catastrophic climate change as an irreversible but non-existential risk; e.g. Cropper 1976; Ulph and Ulph 1997; Tsur and Zemel 2006; Besley and Dixit 2019). This literature, including recent attempts to insert catastrophic risks (half-heartedly) into the IAMs, is reviewed in Tsur and Zemel (2021). Many of these papers are rigorously theoretical, and their conclusions are almost always limited. They typically conclude, rather uselessly, that models incorporating catastrophic risks suggest more investment in mitigation than models ignoring them.

Possibly the best-known exception to the lack of relevance of conventional economic research for the longtermist agenda, specifically as it relates to catastrophic risk, is Weitzman (2009). In his article, Weitzman sets out the 'dismal theorem' (DT) which posits that with catastrophic risks, standard utility-maximizing frameworks cannot be satisfactorily used (see also Nordhaus 2011). In a modest statement atypical for an economics paper, Weitzman says: 'I simply do not know the full answers to the extraordinarily wide range of legitimate questions that DT raises. I don't think anyone does. But I also don't think that such questions can be allowed in good conscience to be simply brushed aside by arguing, in

effect, that when probabilities are small and imprecise, then they should be set precisely to 0' (Weitzman, 2009, p. 13).

4.3 Artificial intelligence

As with climate change, economists obviously cannot be expected to contribute that much to computer-science research on artificial intelligence (AI). But economists obviously ought to be interested in the social consequences of AI, in the design of institutions and incentives that shape AI research, and in how policy can be adjusted to make sure outcomes are beneficial or at the very least benign, even in the long term.

Economists have written many papers discussing AI, but most of this research has a short-term focus, consisting of backwards-looking or short-term forwards-looking analyses of automation, technological unemployment, and inequality (e.g. Autor 2015; Mokyr, Vickers, and Ziebarth 2015). What of the key possibility of core interest to longtermists: an 'intelligence explosion' generated by recursively self-improving AI leading to a transformation of economic growth, likely sometime in the next century (Karnofsky 2022)?

The short-term focus of the existing literature on AI is a natural consequence of the hardness biases we have outlined. To satisfy prevailing methodological standards, empirical work must rely on historical data (or historically informed parameterization), usually covering only a few decades; it is therefore difficult to draw out implications for future developments in AI that may look very different from the preceding waves of automation. The profession is averse to more forward-looking empirical work that, for example, would draw on expert predictions or use existing evidence to generate forecasts about future trajectories of AI research and their impacts.

Theoretical work on AI and automation, which coevolves with the empirical literature as models are developed to explain patterns observed in historical data (such as changes in employment or the wage structure), consequently has a short-term focus. There is a strong appetite for models that explain results observed in the empirical literature, and little appetite for models that rely on controversial assumptions about what future developments in AI might look like.

Economics research that takes seriously the possibility of transformative AI therefore faces steep barriers to publication. Consider research projects on any of the following longtermist topics: analysing which institutional setups best protect social welfare in the event of a non-catastrophic intelligence explosion; solving mechanism design problems for pre-intelligence-explosion AI governance; or improving the quality of AI timeline forecasts by incorporating factors like strategic interactions between different AI developers and governments. Projects in this vein would be highly speculative and not very mathematically tractable (hence not sufficiently 'hard') despite being very important. Moreover, they would likely rely on a mix of recent empirical evidence, subjective forecasts, and theoretical arguments, meaning they would not fit neatly into the standard typology of research. An economist interested in AI would benefit more, professionally, from writing papers that use the standard toolbox of causal inference or structural estimation from historical data, or pure theory, to speak about phenomena like technological unemployment from incremental automation.

As in the climate change literature, the news on this front is not all bad. Over the past few years, economists have become increasingly uneasy about the potential consequences of AI, and abandoned some of their previous complacency (e.g. Acemoglu and Restrepo 2018; Autor 2022). This bodes well for their receptiveness to longtermist perspectives on AI. Moreover, economists have recently started to write speculative papers about AI that take longtermist assumptions seriously (e.g. Aghion, Jones, and Jones 2019; Agrawal, McHale, and Oettl 2019; Cockburn, Henderson, and Stern 2019; Korinek 2019; Korinek and Stiglitz 2019; 2020; 2021; Trammell and Korinek 2023). Many members of the community are hence *aware* of longtermist perspectives on AI and seem to think they should be taken seriously.

Unfortunately, this work has not yet penetrated the leading mainstream economics journals, and does not seem poised to, for the reasons listed above. Many of the aforementioned articles are by some of the most prominent and respected economists within the profession, yet at the time of writing all are either unpublished working papers or book chapters (the latter widely viewed as a publishing venue for papers that are otherwise unpublishable by reputable refereed journals). While individual *economists* have worked on AI from a longtermist perspective, the economics *profession* is not close to doing so due to its systematic hardness bias. Economists must become much more pluralistic before the profession can start to substantively contribute to longtermist AI research.

5 Why economists? Why *academic* economists?

If the kind of research that would be useful for longtermists is radically different from the kinds of research that academic economists have chosen to specialize in, why put the burden of conducting this research on academic economics? Why not suggest that the mantle of conducting useful longtermist research be taken up by other social scientists more used to dealing with qualitative or historical information, or even by economists at research think-tanks and other organizations, who are free of the publishing pressures for promotion in academia and the consequent 'tyranny of the top-5'?

This question really has two parts. First, why economists rather than other social scientists? Second, why should *academic* economists participate rather than leaving this to economists working outside of academia?

The answer to the first question is simple. Producing credible and useful longtermist research will require a cooperative effort across many social sciences. Economics, in particular, can offer a number of important concepts, methods, and stylized facts that are relatively neglected by other disciplines. These include:

- concepts of equilibrium and methods of systematically analysing the disorganized behaviour of many actors;
- a typology of the ways in which the decentralization of choices (i.e. without an optimal, mythical social planner) can lead to perverse consequences—including negative and positive externalities, under-provision of public goods, and negative consequences from information asymmetries or from monopolistic market power;
- practical tools developed by some specialized fields of economics (e.g. mechanism design or non-linear econometrics);

- robust frameworks for thinking about quantitative causal inference;
- ideological blind spots that may be *different* from blind spots in other disciplines, meaning that economists could contribute beneficially to a diversity of thought (for example, economics is the least left-wing social science, and the social science most friendly to markets).

These concepts and methods can contribute, for example, to explanations for why nations underinvest in existential risk reductions, and the corollary of what policies may increase that investment. They can explain why privately funded AI research will produce, possibly inevitably, a fast race towards artificial general intelligence which may be harmful to society's longer-term interests. They can also tell us how scientific grant-making institutions or prediction markets should be designed to incentivize the most useful (and the least dangerous) AI research; they can help us design better long-term mitigation and adaptation strategies for climate change, and so forth.

Importantly, a full longtermist research agenda will demand qualitative, historical, descriptive, predictive, and explanatory work *infused* with these conceptual frameworks and ideas. So a naïve division of labour where, for example, the qualitative work is fully delegated to sociologists and historians, and economists are off in their own corner writing theoretical models of externalities or analysing retrospective data, would not produce the necessary research portfolio. Instead, economists must get their hands dirty with 'softer' kinds of research and collaborate much more closely with researchers from other disciplines.

Second, why should *academic* economists be involved in these efforts, rather than just economists in governments, international organizations, or think-tanks? Those working outside of academia are relatively unconstrained by academic norms and closer to practical decision-makers; wouldn't they be a better fit for this kind of research? Three reasons lead us to argue that both academic and non-academic economists should be involved.

First, academic economists form a large pool of talented individuals motivated (at least in part) by the intrinsic reward of finding convincing answers to important questions. With better incentives, they would provide an additional, and much larger, arsenal to draw on in addition to the efforts of non-academic economists. Second, academia has a unique culture of transparency and intellectual rigour that—while imperfect in many ways—still offers advantages relative to the culture often prevalent in these other institutions. We believe that economics stands out in academia along these dimensions; consider economists' famously intellectually combative and critical (and sometimes, unfortunately cruel) seminars, as well as recent moves towards reproducibility and research transparency in the discipline (moves that are still far from complete but seem to be leading other disciplines).

And finally, academics are also teachers and thus exercise large influence over future generations of talent. Getting academic economists interested and active in longtermist research will therefore create large trickle-down effects to their students, which can percolate for many years (or indeed generations). Anton Korinek, one of the most prolific economists working on this topic, developed a Coursera MOOC (Massive Open Online Course) on AI, with a small section devoted to longtermist issues. This kind of investment in generating interest in the questions posed by longtermist research agendas will not be possible if longtermist economic research stays mostly outside of academia.

6 Alternative explanations

As one reviewer of this chapter noted, there may be alternative explanations for the reluctance of economists to engage with a longer-termist research and policy agendas. One explanation is that economists, like many other people (and many other academics), have not been exposed to or actively reject the basic tenets of longtermism, and therefore do not view it as a plausible focus of research. Another is that economists' research is responsive to the interests of major policy institutions (governments and international institutions like the IMF or World Bank) which are not currently interested in longtermist policymaking.

With respect to the first explanation, we think the main problem is that most economists have not heard of longtermism, not that they actively reject it. As awareness within the general public of longtermism increases, we are optimistic that economists will be receptive. It is true that, as we noted earlier, economists traditionally adopt a conservative approach to the estimation of normative parameters. On the other hand, and partially due to the demands of mathematical tractability, impartialist utilitarian welfare frameworks are practically the only way economists know how to think about normative questions, which bodes well for their receptiveness to longtermism.

Still, we think exposure to and acceptance of longtermist ideas is necessary but not sufficient to compel academic economists to engage in longtermist research. There are many important questions that economists *currently* ignore because of overly strict methodological standards. The addition of new and important longtermist questions to that list will not necessarily change methodological standards by itself. An active push for changes in methodological norms and standards is needed.

With respect to the second explanation, it is true that interest (and funding) from the policy arena—for example from the World Bank or the European Union—plays a role in determining the research focus of economists. However, there is reason to be optimistic about the future direction of these institutions. Economists' rapidly growing interest in climate change in the past couple of years, and the 2020–2021 surge of interest in pandemic economic research, arose in part because of interest from policymakers and funding bodies. Even so, we still view economists' methodological standards as a crucial bottleneck on the ability of funding bodies to entice economists into longtermist research. Most economists would easily choose a top-5 publication over a large research grant, and are hence more responsive to the expectations of their peers than to grantmaking institutions.

7 Practical recommendations

What does our discussion in this chapter imply for what institutions and economic researchers interested in longtermism ought to do? As economists, we are professionally reluctant to offer recommendations that are not backed by cleanly identified causal estimates and contingent on acceptance of a clearly specified welfare framework. But in the spirit of practising what we preach, we will offer one (informal and speculative) recommendation.

We think longtermists should be clear-eyed about the difficulty of attracting career-concerned economists to longtermist research. For example, the Global Priorities Institute runs scholarships and conference programmes for economics PhD students that encourage them to engage in longtermist research. But most PhD programmes implicitly or explicitly

hold up the acquisition of a tenure-track academic job as the ultimate goal for PhD students, and the unfortunate fact is that—given current methodological standards—doing longtermist research is a bad career move for PhD students who have this goal.

We are hence pessimistic about initiatives that target early-career researchers. The problem is that tractable alternatives are not obviously available. One strategy that might be worth considering is instead targeting initiatives at the profession's gatekeepers—the senior economists who edit journals, influence hiring, and thereby shape the direction of research. The economics profession is very centralized and is thus more amenable to 'top-down' than 'bottom-up' change. Senior economists with secure careers have the freedom to pursue research that might not be currently professionally rewarded. In this case, at least, change might be best pursued from above.

References

Acemoglu, D. and Restrepo, P. (2018), 'The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment', in *American Economic Review* 108/6: 1488–1542.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2019), 'Democracy Does Cause Growth', *Journal of Political Economy* 127/1, 47–100.

Aghion, P., Jones, B. F., and Jones, C.I. (2019), 'Artificial Intelligence and Economic Growth,' in A. Agrawal, J. Gans, and A. Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 282–289.

Agrawal, A., McHale, J., and Oettl, A. (2019), 'Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth', in A. Agrawal, J. Gans, and A. Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 282–289.

Akerlof, G. (1970), 'The Market for "Lemons": Quality Uncertainty and the Market Mechanism', *Quarterly Journal of Economics* 84/3: 488–500.

Akerlof, G. (2020), 'Sins of Omission and the Practice of Economics', in *Journal of Economic Literature* 58/2: 405–418.

Andre, P. and Falk, A. (2021), 'What's Worth Knowing? Economists' Opinions About Economics', in IZA Discussion Paper No. 14527.

Autor, D. (2015), 'Why Are There Still So Many Jobs? The History and Future of Workplace Automation', in *Journal of Economic Perspectives* 29/3: 3–30.

Autor, D. (2022), 'The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty,' Working Paper 30074 (NBER Working Paper Series).

Besley, T. and Dixit, A. (2019), 'Environmental Catastrophes and Mitigation Policies in a Multiregion World', in *Proceedings of the National Academy of Sciences* 116/12: 5270–5276.

Burke, M. and Emerick, K. (2016), 'Adaptation to Climate Change: Evidence from US Agriculture', in *American Economic Journal: Economic Policy* 8/3: 106–140.

Chan, G., Carraro, C., Edenhofer, O., Kolstad, C., and Stavins, R. (2016), 'Reforming The IPCC's Assessment of Climate Change Economics', in *Climate Change Economics* 07/1: 1640001.

Chen, S. and Gong, B. (2021), 'Response and Adaptation of Agriculture to Climate Change: Evidence from China', in *Journal of Development Economics* 148: 102557.

Cockburn, I. M., Henderson, R., and Stern, S. (2019), 'The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis', in A. Agrawal, J. Gans, and A. Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 115–146.

Cropper, M. L. (1976), 'Regulating Activities with Catastrophic Environmental Effects', in *Journal of Environmental Economics and Management* 3/1: 1–15.

Dasgupta, P. (2019), *Time and the Generations: Population Ethics for a Diminishing Planet* (Columbia University Press).

Dell, M., Jones, B. F, and Olken, B. A. (2012), 'Temperature Shocks and Economic Growth: Evidence from the Last Half Century', in *American Economic Journal: Macroeconomics* 4/3: 66–95.

Finkelstein, A. and Hendren, N. (2020), 'Welfare Analysis Meets Causal Inference', in *Journal of Economic Perspectives* 34/4: 146–167.

Fourcade, M., Ollion, E., and Algan, Y. (2015), 'The Superiority of Economists', in *Journal of Economic Perspectives* 29/1: 89–114.
Greaves, H. (2017), 'Discounting for Public Policy: A Survey', in *Economics and Philosophy* 33/3: 391–439.
Hassler, J., Krusell, P., and Nycander, J. (2016), 'Climate Policy', in *Economic Policy* 31/87: 503–558.
Heckman, J. and Moktan, S. (2020), 'Publishing and Promotion in Economics: The Tyranny of the Top 5', in *Journal of Economic Literature* 58/2: 419–470.
Howard, P. H. and Sterner, T. (2017), 'Few and Not So Far Between: A Meta-analysis of Climate Damage Estimates', in *Environmental and Resource Economics* 68/1: 197–225.
Hsiang, S., Kopp, R., Jina, A., et al., (2017), 'Estimating Economic Damage from Climate Change in the United States', in *Science* 356/6345: 1362–1369.
Ide, T. (2018), 'Climate War in the Middle East? Drought, the Syrian Civil War and the State of Climate-Conflict Research', in *Current Climate Change Report* 4: 347–354.
Jones. C. I. (2022), 'The End of Economic Growth? Unintended Consequences of a Declining Population', in *American Economic Review* 11/11: 3489–3527.
Kalkuhl, M. and Wenz, L. (2020), 'The Impact of Climate Conditions on Economic Production. Evidence from a Global Panel of Regions', in *Journal of Environmental Economics and Management* 103: 102360.
Karnofsky, H. (2022), 'The "Most Important Century" Blog Post Series', *Cold Takes*, https://www.cold-takes.com/most-important-century/ (accessed 6 January 2025).
Korinek, A. (2019), 'The Rise of Artificially Intelligent Agents', working paper.
Korinek, A. and Stiglitz, J. (2019), 'Artificial Intelligence and Its Implications for Income Distribution and Unemployment', in A. Agrawal, J. Gans, and A. Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 349–390.
Korinek, A. and Stiglitz, J. (2020), 'Steering Technological Progress', working paper. http://rcea.org/wp-content/uploads/2021/04/Future-of-growth/Korinek.pdf (accessed 6 January 2025).
Korinek, A. and Stiglitz, J. (2021), 'Artificial Intelligence, Globalization, and Strategies for Economic Development', Working Paper 28453 (NBER Working Paper Series).
Lamperti, F., Dosi, G., Napoletano, M., Roventini, A., and Sapio, A. (2018), 'Faraway, So Close: Coupled Climate and Economic Dynamics in an Agent-based Integrated Assessment Model', in *Ecological Economics* 150: 315–339.
Millner, A. (2013), 'On Welfare Frameworks and Catastrophic Climate Risks', in *Journal of Environmental Economics and Management* 65/2: 310–325.
Millner, A. and Heal, G. (2023), 'Choosing the Future: Markets, Ethics, and Rapprochement in Social Discounting', in *Journal of Economic Literature* 61/3: 1037–1087.
Mokyr, J., Vickers, C., and Ziebarth, N. L. (2015), 'The History of Technological Anxiety and the Future of Economic Growth: Is This Time Different?', in *Journal of Economic Perspectives* 29/3: 31–50.
Nordhaus, W. D. (1975), 'Can We Control Carbon Dioxide?', IIASA Working Paper 75-63 (International Institute for Applied Systems Analysis).
Nordhaus, W. D. (1992), 'An Optimal Transition Path for Controlling Greenhouse Gases', in *Science* 258/5086: 1315–1319.
Nordhaus W. D. (2011), 'The Economics of Tail Events with an Application to Climate Change', in *Review of Environmental Economics and Policy* 5/2: 240–257.
Noy, I. (2023), 'Economists Are Not Engaged Enough with the IPCC', in *npj Climate Action* 2: 33.
Noy, I. and Uher, T. (2022), 'Four New Horsemen of an Apocalypse? Solar Flares, Super-volcanoes, Pandemics, and Artificial Intelligence', in *Economics of Disasters and Climate Change* 6/2: 393–416.
Ramsey, F. (1928), 'A Mathematical Theory of Saving', in *The Economic Journal* 38/152: 543–559.
Rodrik, D. (2021), 'How Economics and Non-Economists Can Get Along,' *Project Syndicate*, https://www.project-syndicate.org/commentary/economists-other-social-scientists-and-historians-can-get-along-by-dani-rodrik-2021-03 (accessed 6 January 2025).
Romer, P. M. (1990), 'Endogenous Technological Change', in *Journal of Political Economy* 98/5: S71–S102.
Stern, N. (2007), *The Economics of Climate Change: The Stern Review* (Cambridge University Press).
Stern, N. (2013), 'The Structure of Economic Modeling of the Potential Impacts of Climate Change: Grafting Gross Underestimation of Risk onto Already Narrow Science Models', in *Journal of Economic Literature* 51/3: 838–59.
Stern, N. and Stiglitz, J. E. (2022), 'The Economics of Immense Risk, Urgent Action and Radical Change: Towards New Approaches to the Economics of Climate Change', in *The Journal of Economic Methodology* 9/3: 181–216.
Timilsina, G. (2022), 'Carbon Taxes', in *Journal of Economic Literature* 60/4: 1456–1502.

Trammell, P. and Korinek, A. (2023), 'Economic Growth under Transformative AI: A Guide to the Vast Range of Possibilities for Output Growth, Wages, and the Labor Share, GPI Working Paper No. 8-2020 (Global Priorities Institute, Oxford University).
Tsur, Y. and Zemel, A. (2006), 'Welfare Measurement under Threats of Environmental Catastrophes', in *Journal of Environmental Economics and Management* 52/1: 421–429.
Tsur, Y. and Zemel, A. (2021), 'Resource Management under Catastrophic Threats', in *Annual Review of Resource Economics* 13/1: 403–425.
Ulph, A. and Ulph, D. (1997), 'Global Warming, Irreversibility and Learning', in *The Economic Journal* 107/442: 636–650.
Weitzman, M. L. (2009), 'On Modeling and Interpreting the Economics of Catastrophic Climate Change', in *Review of Economics and Statistics* 91/1: 1–19.

30

The Intuitive Appeal of Legal Protection for Future Generations

Eric Martínez and Christoph Winter

1 Introduction

Over the past several years, there has been a growing interest in protecting future generations from extreme risks associated with climate change, pandemics, artificial intelligence, and other potential threats. This interest has materialized in the form of advocacy efforts (Schoch-Spana et al. 2017; Yassif 2017; Setzer and Vanhala 2019; Bogojević 2020; Bliss 2022), as well as philosophical theories. The theories associated with the view that one should be particularly concerned with ensuring that the long-run future goes well have been referred to as *longtermism* (MacAskill 2022). In the context of law, these theories form the basis for *legal longtermism*, associated with the view that law and legal institutions ought to protect the far future (see Martínez and Winter 2021a; Winter et al. 2021).[1]

Given the recency of this work, as well as the apparent lack of protection afforded to future generations under current legal systems, one implicit assumption surrounding this work has been that the principles underlying longtermism are not intuitive or widely accepted. Even one of longtermism's pioneers, Toby Ord (2020: 7–8), characterizes longtermism as deeply counterintuitive, the sort of theory that only a philosopher could endorse, after years of slow, careful reflection:

> I have not always been focused on protecting our longterm future, coming to the topic only reluctantly . . . Since there is so much work to be done to fix the needless suffering in our present, I was slow to turn to the future. It was so much less visceral; so much more abstract. Could it really be as urgent a problem as suffering now?

In this chapter, we present recent empirical work suggesting that the basic principles underlying legal longtermism are intuitive once people are made aware of them.[2] In particular, evidence suggests that most people across major demographic subgroups believe that (i) law *should* protect the long-term future much more than it currently does; and (ii) law *can* predictably and feasibly protect the long-term future. We also review ideas associated

[1] For the purposes of this chapter, we are proceeding with this rather broad definition. Indeed, one could further differentiate between *philosophical-level legal longtermism*, that is, the view that the law should be particularly concerned with ensuring that the long-term future goes well independent of existing legal doctrine, and *doctrinal-level legal longtermism*, the view that law should be particularly concerned with ensuring that the long-term future goes well according to the best interpretation of existing legal doctrine.

[2] For the purposes of this chapter, an 'intuitive' idea refers to one that is readily endorsed by the majority of surveyed people across major demographic subgroups without the need for explicit attempts at persuasion.

Eric Martínez and Christoph Winter, *The Intuitive Appeal of Legal Protection for Future Generations* In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press.
© Eric Martínez and Christoph Winter 2025. DOI: 10.1093/9780191979972.003.0030

with stronger forms of legal longtermism that are less intuitive—such as the claim that the law, other things being equal, should protect future people *to the same degree as present people*.

Uncovering the intuitive appeal of legal longtermism has implications both for legal theory and for practitioners of these ideas. On the legal theory front, many plausible theories treat consensus, whether derived from experts or laypeople, as indicative or even constitutive of legal truth (Hart 1961; Baude 2015; Baude and Sachs 2019).[3] Accordingly, that legal experts and laypeople largely agree that law should take seriously the interests or well-being of future people based on the best or ordinary understanding of legal doctrine supports the conclusion that law should take those interests seriously. On the practical front, the empirical evidence suggests that, since many people find these principles appealing once they become aware of them, resources may be best spent raising broad awareness of these principles (as opposed to explicit persuasion attempts) and analyzing how to transform these broad principles into the right actions and policies.

The rest of the chapter proceeds as follows. Section 2 discusses certain *a priori* reasons to expect longtermism and legal longtermism to be intuitive or counterintuitive. Section 3 presents empirical data from recent studies documenting the intuitive appeal of legal longtermism. Section 4 discusses the limitations and implications of these results for longtermist theory and practice.

2 Intuitions about the intuitive appeal of (legal) longtermism

This section discusses certain *a priori* reasons to expect that (legal) longtermism might or might not be intuitively appealing to the general public.

2.1 The supposed counterintuitiveness of legal longtermism

Despite the development of philosophical longtermism and legal longtermism as academic theories, there are some reasons to expect that the principles underlying both may not be intuitive or widely accepted.[4]

First, longtermism has been described by adherents and detractors alike as being counterintuitive. For instance, Whittlestone (2022) stated that some might be skeptical because

[3] E.g. Baude and Sachs (2019: 1464) explain, 'positivism grounds law in social practice and consensus'. Although this consensus is often understood as expert consensus among legal officials or academics, legal doctrine is in other respects explicitly grounded in the notion of lay consensus. For instance, according to the ordinary meaning doctrine, considered to be the most 'fundamental principle of legal interpretation' (Slocum 2019), the words of a legal document are generally to be interpreted according to how they are ordinarily understood by laypeople. For examples of jurists more broadly arguing that lay consensus is or ought to be informative of legal doctrine, see e.g. Tobia (2021: 86) ('The broader conclusion is that cognitive scientists can make significant progress in understanding legal cognition—and law itself—by studying the ordinary cognition of people with no special legal training.'); Martínez and Tobia (2023: 182) (stating that one might conclude that 'valuations of what primary purpose law should serve, or what considerations should inform judgments of reasonableness, should be determined by laypeople as opposed to legal experts'). See also Tobia (2022); Martínez and Winter (2024).

[4] Note that in this chapter we do not consider whether the line of reasoning is sound or valid, but instead what might be intuitive to some. As we later show, recent empirical evidence suggests that the basic abstract principles of longtermism and legal longtermism are, in fact, fairly intuitive.

'prioritising the long-term future . . . is a counterintuitive way of doing good'. Kannan (2021) likewise listed 'longtermism being counterintuitive' as the first of many 'common standard counter-arguments' to longtermism, and Ord (2018) described the longtermist cause of existential risk as being often framed as 'this really counterintuitive idea'.[5]

A second reason relates to the size and lack of intuitive appeal of the movement which longtermism grew out of. Longtermism grew out of effective altruism (EA), a philosophical and social movement that advocates 'using evidence and reason to figure out how to benefit others as much as possible, and taking action on that basis' (MacAskill 2017: 2).[6] The term 'effective altruism' was coined in 2011, and the movement remains small to this day; according to recent surveys, there are fewer than 12,000 EA community members, who represent less than 1/500,000 of the global population (Moss 2020; 2021).[7] Effective altruism has been described by its adherents (e.g. MacAskill 2015), detractors (e.g. Wu 2022), and neutral commentators (e.g. Lewis-Kraus 2022)[8] alike as taking a counterintuitive approach to doing good. In other words, the fact that longtermism was developed by a niche community with a counterintuitive approach to doing good might indicate that longtermism itself is counterintuitive.

A third reason relates to the recency of philosophical longtermism as a formal philosophy. Philosophy is one of the oldest academic disciplines, and for millennia scholars have formalized, developed, and documented moral intuitions into formalized theories. One might expect that, on that time frame, moral views even somewhat intuitive would have already been developed into formalized moral theory. However, although Sidgwick (1907) and other philosophers appear to have thought carefully about ideas consistent with longtermism in the 19th century or even earlier,[9] longtermism and legal longtermism did not exist as formalized theories until well into the 21st century. Thus, the recency of longtermism within philosophy may lead one to expect that it is not as intuitive as other theories that were developed centuries or even millennia earlier and which enjoy mainstream acceptance among professional philosophers, such as Kantian deontology, Aristotelian virtue ethics, or even Benthamite utilitarianism.

In the context of legal longtermism specifically, the fourth and perhaps most salient reason to expect longtermism to be counterintuitive is the sheer lack of *de facto* legal protection provided to future generations. For example, although there have been various dissenting voices over time advocating for 'the future of man' via environmental protections

[5] Some may object to these quotes as referring to certain implications of longtermism as opposed to the basic abstract principles. While this may be the case, and we will discuss some less intuitive aspects of longtermism's implications later, we believe these quotes are most naturally understood as being about longtermism in general.

[6] Though also note that many others were developing some of the foundational ideas of longtermism long before effective altruism existed or 'longtermism' was coined. See, for instance, Russell and Einstein (1955), von Neumann (1955), Baier (1981), Parfit (1984), Sagan (1994), UNESCO (1997). For an exploration of the history of thinking about human extinction, see Moynihan (2020).

[7] Though, as we will describe later, one reason for this fact might be that most people simply have not yet come across EA. See Caviola, Morrissey, and Lewis (2022).

[8] 'Effective altruists have lashed themselves to the mast of a certain kind of logical rigor, refusing to look away when it leads them to counterintuitive, bewildering, or even seemingly repugnant conclusions' (Lewis-Kraus 2022). See also Pellegrino (2017: 44), arguing that certain tenets of effective altruism require effective altruists to 'accept some strongly counterintuitive judgments'.

[9] See fn. 6. See further MacAskill's (2022: 76) discussion of the Mohists, a group of adherents to a consequentialist philosophy whose teachings in the 5th century BC were said to 'fill the world'. MacAskill notes that the silencing and oppression of Mohism may have contributed to a 'value lock-in' of values less favorable to longtermism, such as Confucianism (2022: 97–8).

and nuclear disarmament (Mansfield 1955; Russell and Einstein 1955), past and present legal systems have failed to grant future generations democratic representation in the legislature,[10] standing to bring forth a lawsuit in the judiciary,[11] and serious consideration in cost-benefit analyses in the executive.[12] Were legal longtermism intuitive, this line of reasoning holds, surely democratic legal systems would have granted future generations some form of substantial protection by now.

2.2 The arguable intuitiveness of legal longtermism

Nevertheless, one may also point to reasons legal longtermism could be expected to be intuitive or commonsensical. First among them, philosophical longtermism appears to be compatible with a wide range of moral philosophies, including deontology (Baier 1981), virtue ethics (Brand 2000; Schell 2000; Ord 2020),[13] and consequentialism (Bourget and Chalmers 2021; Martínez and Winter n.d.).[14]

A second reason is that, despite the fact that legal systems do not afford *de facto* legal protection to future generations, many legal systems are beginning to provide *de jure* legal protection to future generations independent of other issues, such as climate change.[15] For example, recent work by Araújo and Koessler (2021) found that constitutions referencing future generations now comprise roughly one-third of all the constitutions in force. Although most of these reference future generations alongside or in the context of environmental protection (62%) and natural resources (35%), some constitutions (22%) mention future generations *stricto sensu*—that is, by themselves without another theme mentioned. The past few decades have also seen a handful of institutions and government offices designed to consider the interests of future generations, such as the Future Generations Commissioner of Wales, although to date these have focused primarily on environmental

[10] While not possible directly via the right to vote, it could be done indirectly via representation in the legislature, for example (González-Ricoy and Gosseries 2016; John and MacAskill 2022).

[11] Although standing requirements can vary widely across jurisdictions, the vast majority of jurisdictions have not explicitly extended the doctrine of *locus standi* to future generations. See e.g. Bogojević (2020), discussing the challenges and failures of extending the doctrine of standing in climate law cases. But see *Minors Oposa* (Supreme Court of the Philippines 1993, stating, 'We find no difficulty in ruling that [petitioners] can, for themselves, for others of their generation and for the succeeding generations, file a class suit.'); Sahoutara (2016) and *Rabab Ali v. Pakistan* (Petition to Supreme Court of Pakistan 2016: paras. 1, 6–7, 31, xviii) (granting standing to petitioner challenging various government actions related to Thar coal, on behalf of present and future generations).

[12] For an overview of discount rates, see Zhuang et al. (2007).

[13] For example, one might argue from a deontological perspective that we owe a duty to future generations, independent of what a consequentialist calculus might demand. From a virtue ethics perspective, one might argue that it is a virtue to act in such a way that protects future generations by exercising 'civilizational virtues' such as patience, self-discipline, benevolence, and taking responsibility for our actions (Gaba 1999: 283–7; Ord 2020; Winter et al. 2021).

[14] In Bourget and Chalmers' (2021) survey of professional philosophers, the distribution of participants' normative ethical beliefs was as follows: 37.2% endorsed virtue ethics, 32.1% endorsed deontology, and 30.6% endorsed consequentialism. In Martínez and Winters's (n.d.) survey of legal academics, the distribution was as follows: 49.5% endorsed deontology, 41.1% endorsed consequentialism, and 57.2% endorsed virtue ethics (note that Martínez and Winters's format allowed for participants to endorse more than one theory).

[15] Note that for the purposes of this chapter, *de jure* protection refers to protection that is officially provided by written law (so-called 'law in books') but not necessarily recognized or enforced in practice by the courts (so-called 'law in action'), whereas *de facto* protection refers to protection that is actually recognized in practice by the courts.

protection and sustainability (see e.g. Jones, O'Brien, and Ryan 2018; Olawuyi 2021; Viña and Bueta 2021). These efforts to protect future generations suggest that lawmakers, and the constituents and groups who inform their policies, may find the principles of legal longtermism intuitively appealing.

3 Legal longtermism's empirical intuitive appeal

Recent empirical work appears to support the idea that, at least on an abstract level, people agree with the principles of legal longtermism. In this section, we present three sets of findings in support of this claim, showing that: (i) legal experts and laypeople alike believe that the law *should* protect the long-term future much more than it currently does, in subsection 3.1; (ii) legal experts believe that the law *can* predictably and feasibly protect the long-term future, in subsection 3.2; and (iii) these beliefs hold true across major demographic subgroups and across cultures, in subsection 3.3.

These findings draw on four recent empirical studies:

1. Global Law Professor Survey of over 500 legal academics from leading universities around the English-speaking world, specifically in Australia, Bangladesh, Canada, India, New Zealand, South Africa, and the United Kingdom (Martínez and Winter 2021a);
2. U.S. Law Professor Survey of over 600 United States law professors[16] (Martínez and Tobia 2023);
3. U.S. Layperson Survey of over 1,000 lay adults in the United States (Martínez and Winter 2021b); and
4. Global Layperson Survey of roughly 3,000 lay adults across 10 countries—Australia, Canada, Chile, Japan, Mexico, Spain, South Africa, South Korea, the United Kingdom, and the United States (Martínez and Winter forthcoming).

3.1 The law should do more to protect the far future

One source of evidence comes from a pair of recent empirical studies. Both the Global Law Professor Survey and the U.S. Layperson Survey asked participants about their beliefs regarding the current and desired level of legal protection afforded to future generations and other groups, such as present humans, non-human animals, the environment, and artificial intelligence. The two prompts were presented to participants as follows:

1. On a scale of 0–100, how much **does** your country's legal system protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?

[16] Note that United States law professors refers to professors based in the United States of America (as opposed to, for example, professors that necessarily specialize in United States law).

2. On a scale of 0–100, how much **should** your country's legal system protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?

With regard to the rating scale, 0 represented 'not at all' and 100 represented 'as much as possible'.[17] Participants were asked to rate the following groups:

Humans inside the jurisdiction
Humans outside the jurisdiction
Non-human animals
Environment (e.g. rivers, trees, or nature itself)
Sentient artificial intelligence (assuming its existence)
Humans living now
Humans living in the near future (0–25 years from now)
Humans living in the medium future (25–100 years from now)
Humans living in the far future (100+ years from now)

Some of the main results of these studies are visualized in Figures 30.1 and 30.2.[18] The main takeaways are threefold.

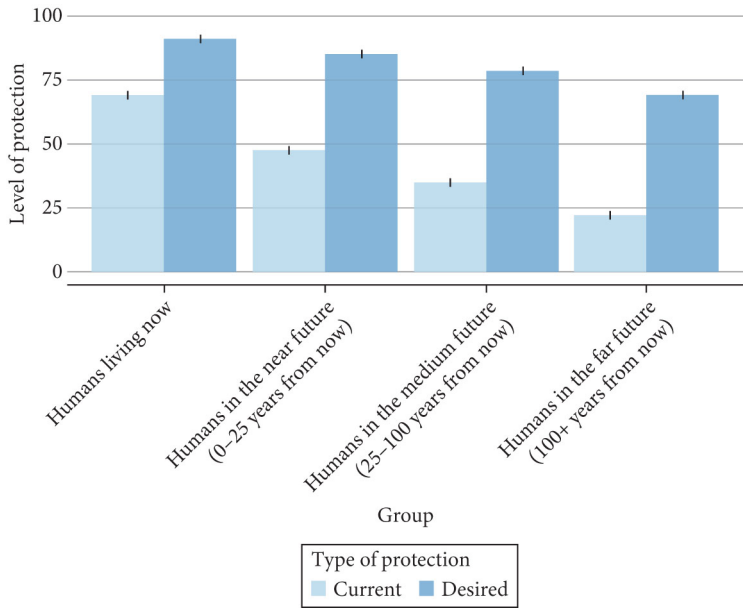


Figure 30.1 Current vs. desired level of legal protection for present and future humans (Global Law Professor Survey) (Martinez and Winter 2021a).

[17] For simplicity, we refer to these ratings as the current and desired levels of protection and compare them across different groups. Note that the max endpoint, 'as much as possible', could be interpreted as being different across groups, if a respondent believes that a different amount of legal protection is possible for each group. Thus, a higher (or lower) score for one group may not in fact reflect a higher (or lower) level of current or desired protection compared to another group.

[18] Note that in all figures, error bars reflect 95% bootstrapped confidence intervals.

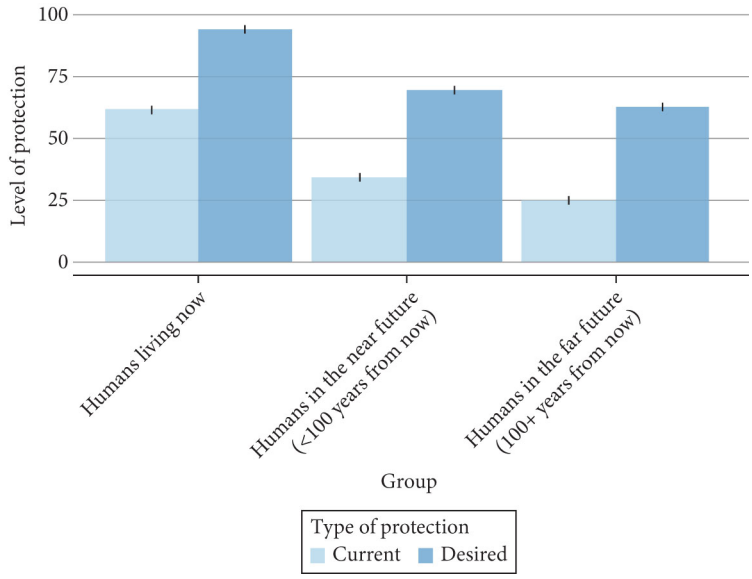


Figure 30.2 Current vs. desired level of legal protection for present and future humans (U.S. Layperson Survey) (Martinez and Winter 2021b).

First, the desired level of protection to future generations was roughly two-and-a-half to three times higher than the perceived current level of protection. Second, the desired level of protection for those living in the far future was roughly the same as the perceived current level of protection being afforded to humans living in the present. Third, the gap between the desired and current level of protection is (i) significantly higher for future generations than the present generation; (ii) significantly higher for humans living in the far future than the near future; and (iii) higher for humans living in the far future than any other group surveyed on, including non-human animals, the environment, artificial intelligence, and humans outside the jurisdiction.[19] Taken together, these results support the claim that legal experts and laypeople alike believe the law should do much more to protect humans living in the long-term future, even when their interests are weighed against other neglected entities and issues.[20]

Empirical work also suggests that future generations are deserving of not only general legal protection, but fundamental access to the legal system, such as via legal personhood. In the U.S. Law Professor Survey, Martínez and Tobia (2023) asked participants their beliefs regarding several dozen legal theory issues, including personhood. The personhood prompt was presented as follows:

> Insofar as domestic law should protect the rights, interests, and/or well-being of 'persons', which of the following categories includes at least some 'persons'?

[19] Note that 'significantly' here (and elsewhere when discussing the results) refers to statistical significance. For details regarding the statistical analyses conducted for the survey, refer to Martínez and Winter (2021b).

[20] Additionally, the fact that the desired level of protection for those living in the far future (even taking the lower bound of the 95% confidence interval) was non-zero alone provides evidence in favor of the intuitive appeal of the principle that future people count and deserve legal consideration. Moreover, the fact that the *current* level of protection for those living in the far future was also non-zero provides further evidence in favor of the intuitive appeal of the principle that that law can affect the lives of future people in a positive way.

Participants were asked about the following categories:

Humans in the legal jurisdiction
Humans outside the legal jurisdiction
Corporations
Unions
Non-human animals
Artificially intelligent beings
Humans who do not yet exist, but will be born in the next 50 years
Humans who will only exist in the very distant future

For each category, participants could either 'Accept', 'Lean towards', 'Lean against', or 'Reject' personhood, or they could choose from a number of 'other' options as an explanation for why they could not provide a rating (e.g. 'it depends', 'insufficient knowledge', or 'no fact of the matter').

A majority of U.S. law professors surveyed (53.8%) leaned towards or accepted personhood for humans who will be born in the next 50 years, while a substantial minority (34.5%) leaned towards or accepted personhood for humans who will only exist in the very distant future.

An extension of this work, the U.S. Layperson Survey asked U.S. adults the same prompt with slight differences in the categories surveyed on—instead of 'humans who do not yet exist, but will be born in the next 50 years' and 'humans who will only exist in the very distant future', participants were asked about 'humans living in the near future (<100 years from now)' and 'humans living in the far future (100+ years from now)'. Participants favored personhood for future persons: 64.09% leaned towards or accepted personhood for humans living in the near future, and 61.75% leaned towards or accepted personhood for humans living in the far future.

Given that future humans in general are not currently granted personhood in any fashion, these findings further support the claim that experts and laypeople believe the law should protect the future more than it does currently.

3.2 The law can protect the far future

Other empirical evidence supports the premise that law *can* help those living in the future. The Global Law Professor Survey asked participants whether they believed there are feasible, predictable mechanisms through which the law can affect the long-term future. Participants were asked about the long-term future (defined as at least 100 years from now) as well as the very long-term future (defined as at least 1,000 years from now). Figure 30.3 visualizes the main results. For both time periods, significantly more law professors agreed than disagreed with the claim that law can predictably and feasibly influence the future. The vast majority (74.5%) of participants agreed with respect to the long-term future, while a plurality (40.9%) agreed with respect to the very long-term future—both striking results given longtermist concerns about cluelessness and washing out (Greaves and MacAskill 2021; Thorstad 2021).[21]

[21] Note that 'cluelessness' in longtermist literature refers to the concern that it is impossible to calculate the expected value for long-term interventions. The 'washing-out' hypothesis refers to the concern that it is impossible to influence the far future, given the likelihood that the effects of one's actions and policies decay over time, making the effects on the near-term outweigh any on the long-term. Note also that participants' estimates may also reflect unfamiliarity with the longtermist concepts of cluelessness and washing out, as well as lack of expertise in forecasting discussed by Martínez and Winter (2021a: 38, note 82). For further discussion of cluelessness, see Greaves (2016). For further discussion of the washing-out hypothesis, see Greaves and MacAskill (2021).

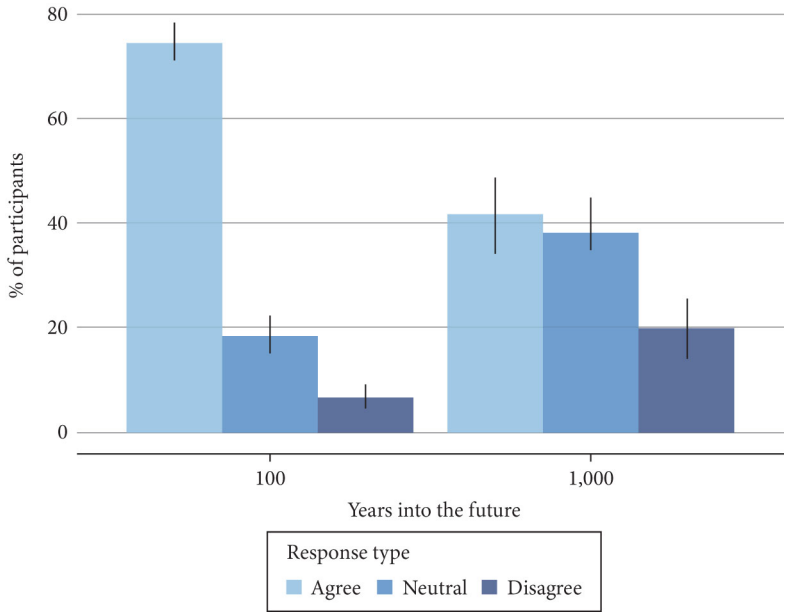


Figure 30.3 Participant responses to the prompt that there are feasible, predictable legal mechanisms for influencing the future (Global Law Professor Survey) (Martinez and Winter 2021a).

Participants responded similarly when asked about specific areas of law, such as constitutional, criminal, and environmental law, as well as when asked about specific risks, such as those resulting from artificial intelligence, climate change, and biorisk, suggesting that legal academics view law as a multifaceted and versatile tool to influence the long-term future.

Similar to the normative prompts discussed in the previous section, several questions in the Global Law Professor Survey asked about more concrete legal mechanisms within existing legal doctrine, such as standing. Participants were asked with respect to several groups whether they considered there to be 'a reasonable legal basis for being granted standing to bring forth a lawsuit (*locus standi*) in at least some possible cases?'

More than two-thirds (67.74%) of legal experts endorsed there being a reasonable legal basis for granting standing to humans living in the near future (understood as up to 100 years from now), while a slight majority (51.16%) endorsed the proposition with regard to humans living in the far future (understood as 100+ years from now). Insofar as legal experts believe standing to be an effective mechanism for influencing the long-term future, this finding supports the claim that experts believe existing law can influence the long-term future.

Finally, the Global Law Professor Survey asked participants to rate how much protection certain constitutional mechanisms, if incorporated into their country's constitution, would provide to future generations. Participants rated the level of protection on a scale of 1 to 7, with 1 representing 'none at all', 4 representing 'some', and 7 representing 'very much'. Participants were asked to rate the following mechanisms:

1. protection against discrimination towards future generations;
2. commitment to spend 1% of GDP towards protection against existential risk (such as those posed by runaway climate change, artificial intelligence, or pandemics);
3. provision granting standing (*locus standi*) to future generations;
4. commission or ombudsperson to oversee the protection of future generations;
5. state goal to protect future generations.

Although some mechanisms were rated as higher than others, for each of these mechanisms,[22] the mean rating for level of protection was above a 4 ('some'), further suggesting that law professors believe specific mechanisms could be implemented to offer at least some protection to the long-term future.[23]

3.3. On the robustness of the intuition that law can and should do more to protect the long-term future

With regard to the third set of findings, the studies revealed that these beliefs are held, not only by legal experts and laypeople as a whole, but also across several demographic subgroups. The findings of both the Global Law Professor Survey and the U.S. Layperson Survey were robust to differences in gender, age, country of origin, type of legal training, and political affiliation.[24] For example, the breakdown of responses to desired vs current level of legal protection among liberal (left-leaning) and conservative (right-leaning) lay adults is visualized in Figure 30.4, which shows that main trends identified with regard to laypeople as a whole were also observed across the political spectrum.

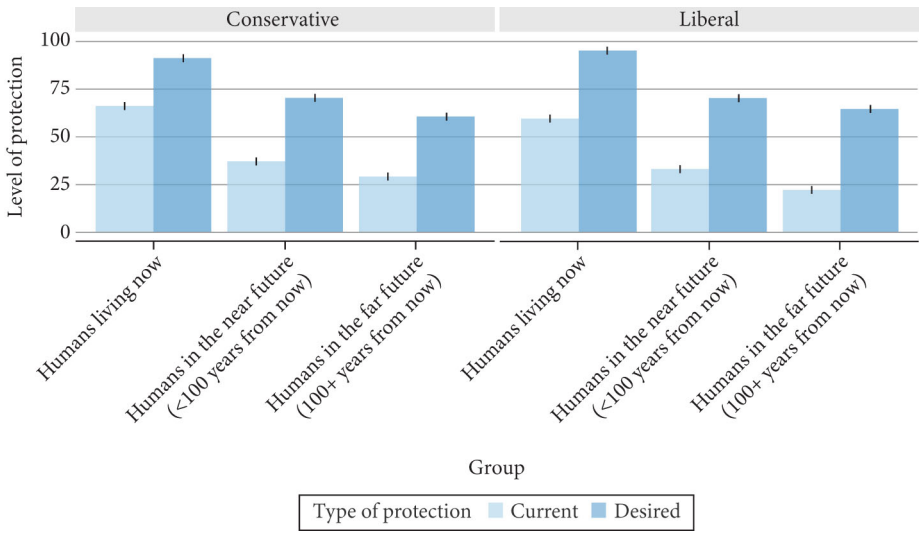


Figure 30.4 Current vs desired protection for present and future humans by political affiliation (U.S. Layperson Survey) (Martinez and Winter 2021b).

[22] The mechanism that was rated as granting the most protection was a commitment to spend 1% of GDP towards protection against existential risk (4.76; 95% CI: 4.50 to 5.05), followed by protection against discrimination towards future generations (4.62; 95% CI: 4.34 to 4.91), provision granting standing to future generations (4.22; 95% CI: 3.92 to 4.53), state goal to protect future generations (4.15; 95% CI: 3.88 to 4.43), and commission or ombudsperson to oversee the protection of future generations (4.13; 95% CI: 3.83 to 4.42).

[23] It's also worth noting that the fact that the law professors gave similar responses for each mechanism may indicate a high level of uncertainty among law professors regarding which mechanism would provide the strongest type of protection.

[24] While Martínez and Winter (2021b) do not describe the demographic findings related to future generations in the U.S. Layperson Survey, the datasets for the study can be found on OSF at https://osf.io/2hfx6/?view_only=25d06cdb33004cfa88ac76ae4a28a5b6.

Additionally, a fourth study indicates that at least some of these beliefs are held not only by people across the anglosphere but across cultures as well. In the Global Layperson Survey, which covered 10 countries, Martínez and Winter (forthcoming) asked participants about their beliefs regarding the current and desired level of legal protection afforded to future generations and other groups. Unlike the U.S. Layperson Survey, the Global Layperson Survey asked participants about protection at both the national level and the international level. At the national level, the wording of the prompts was as follows:

1. On a scale of 0–100, how much **does** your country's legal system protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?
2. On a scale of 0–100, how much should your country's legal system protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?

At the international level, the wording of the prompts was as follows:

3. On a scale of 0–100, how much do international organizations (such as the United Nations) protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?
4. On a scale of 0–100, how much should international organizations (such as the United Nations) protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?

For each prompt, participants were asked about the same groups as in the U.S. Layperson Survey. Results for the national-level prompts are visualized in Figure 30.5.

The results were convergent with those of the U.S. Layperson Survey. In each country: (i) the desired level of protection for humans living in the near and far future was higher than the perceived current level of protection afforded to them; and (ii) the difference between the desired and current level of protection was disproportionately higher for humans living in the near and far future than for other groups as a whole.

As with the U.S. Layperson Survey, the Global Layperson Survey asked participants whether their country's legal system should grant personhood and standing to at least some subset of humans living in the near and far future. In each of the 10 countries, the majority of participants endorsed personhood for at least some individuals within the categories of 'humans living in the near future' and 'humans living in the far future', indicating that granting personhood to future humans is widely supported cross-culturally. With regard to standing, both the majority of participants and the majority of each of the 10 countries endorsed standing both for humans living in the near future (58.2%) and for humans living in the far future (55.6%), indicating that granting standing to humans living in the future likewise enjoys substantial cross-cultural support outside the English-speaking world.

Finally, the Global Layperson Survey also asked whether people believed that the welfare (broadly understood as the rights, interests, and/or well-being) of future generations should ever outweigh that of the present generation. The majority across all participants, and the majority of countries in the sample (7 of 10), endorsed the proposition that there

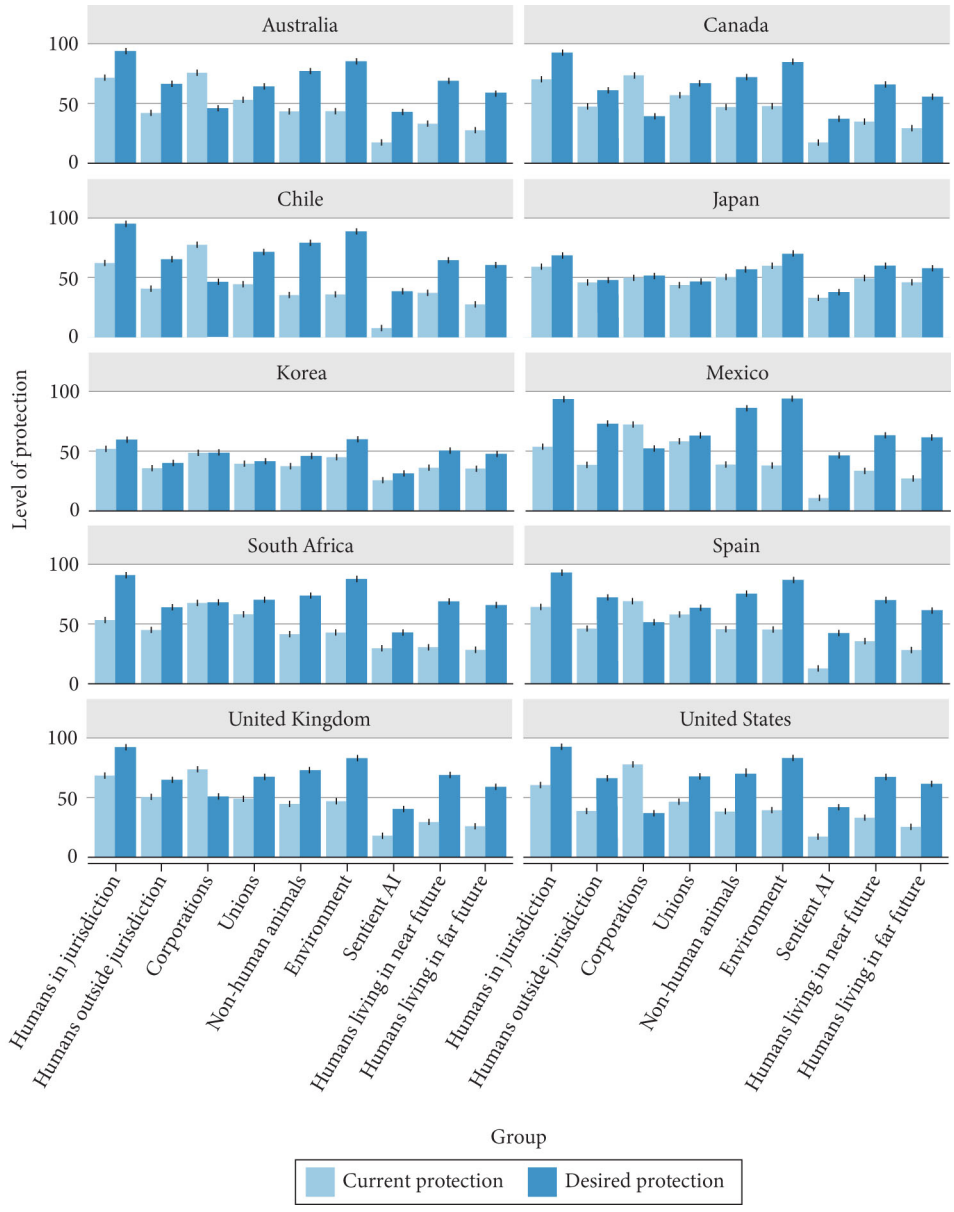


Figure 30.5 Current vs. desired national legal protection of far-future humans and other groups as judged by participants across countries (Global Layperson Survey) (Martinez and Winter forthcoming).

are at least some possible scenarios in which the welfare of future people should outweigh that of present people, both in the context of national law-making (51.4%) and in the context of international law-making (54.1%). This suggests that there is widespread intuitive appeal, not only for increasing legal protection to future generations, but for prioritizing this protection over other groups in certain cases, both at the national and the international level.

4 Discussion

In this section we discuss objections to and implications of the results presented in section 3. First, we present and discuss some evidence suggesting that some of the strongest forms of legal longtermism may not be as intuitive as the basic principles (subsection 4.1). Second, we discuss some legal-philosophical and legal-doctrinal implications of our results for the legal system and for the validity of legal longtermism (subsection 4.2). Third, we discuss the implications of our results for applied longtermism (subsection 4.3).

4.1 Limitations of legal longtermism's intuitive appeal

Despite evidence that people across major demographic groups believe that the law can and should protect the long-term future more than it does currently, there is also evidence that they do not prefer equal protection for the near and far future. Although people may intuitively accept that future people count and law can protect them, people may *not* intuitively accept (i) that far-future people count *just as much* as near-future or present people, nor (ii) that law can predictably and feasibly influence the *very* long-term future, nor (iii) that there could be many future people. We discuss each of these in turn, then describe the potential for further disagreement regarding how best to influence the long-term future.

With respect to whether far-future people count just as much as near-future or present people, reconsider the Global Law Professor Survey, U.S. Layperson Survey, and Global Layperson Survey, all of which asked participants about their beliefs regarding the current and desired level of legal protection afforded to future generations and other groups. In all studies, participants rated the desired level of legal protection for humans living in the far future as two-and-a-half to three times greater than what is currently afforded to them, and in the Global Layperson Survey the majority of participants endorsed the proposition that there are at least some possible scenarios in which the welfare of future people should outweigh that of present people; however, in all studies, the desired level of legal protection for humans living in the far future was still significantly lower than for humans living in the present. This indicates that, although legal experts and laypeople believe that future generations should be legally protected to a greater degree than they are currently, and should even be prioritized over present generations in at least some scenarios, they by and large also believe that future generations should not be legally protected to the same degree as humans living now.[25]

Consider also the U.S. Law Professor Survey, in which participants were asked to endorse or reject personhood for future humans and other groups. Although the majority of participants endorsed personhood for humans who do not yet exist but will be born within the next 50 years, the majority of participants did not endorse personhood for humans

[25] Note also that the max endpoint of the scale (i.e. 100 out of 100) of the prompt was 'as much as possible', indicating that the lower desired level of legal protection for future generations was not a result of participants' simply believing it to be less tractable to protect future generations through the law.

who will only exist in the distant future. Even more participants endorsed personhood for present humans living within the jurisdiction. Taken together, these indicate that U.S. law professors do not believe that the law should provide humans living in the future the same legal status as humans living in the present.

With respect to whether the law can protect the far future, consider again the Global Law Professor Survey. Although the vast majority of law professors agreed with the prompt that there were predictable, feasible mechanisms through which the law could influence the long-term future (understood as at least 100 years from now), only a plurality agreed with the prompt regarding the *very* long-term future (understood as at least 1,000 years from now). Similar results were observed with respect to individual areas of law (e.g. constitutional law, criminal law, environmental law), indicating that most law professors are less confident regarding law's ability today to predictably and feasibly positively influence the future a thousand or more years from now.[26]

Finally, in addition to whether future people count and whether law can protect them, it remains largely an open question whether people intuitively accept the idea that many will exist in the future. As mentioned before, the fact that legal experts are confident that there are feasible, predictable legal mechanisms through which the law can influence the long-term future may imply that legal experts believe there will be people in the future for law to influence. However, there remains no direct evidence of *how many* people are intuited to exist in the far future by people living in the present. Given that stronger versions of longtermism assume high numbers of future people in expectation,[27] and it remains unclear whether people intuitively accept this threshold, it therefore remains unclear whether people endorse the abstract principles underlying stronger forms of longtermism.[28]

Furthermore, those who agree with the basic premises of legal longtermism may still disagree with proponents of longtermism on how best to influence the long-term future. For example, whereas reducing risk from artificial intelligence is typically seen as one of the leading cause areas within longtermism (Karnofsky 2016; Greaves and MacAskill 2021: 13–15; Hilton 2022; Open Philanthropy 2023) and legal longtermism (Winter et al. 2021), fewer than 50% of all participants in the Global Law Professor survey agreed that law could feasibly and predictably influence the long-term future in the context of

[26] For further discussion of the typical time frames considered by longtermists and strong longtermists, see Greaves and MacAskill (2021).

[27] Greaves and MacAskill (2021: 10) consider different scenarios in estimating the number of future beings and conclude that 'any reasonable estimate of the expected number of future beings is at least 10^{24}', with upper estimates of 10^{36} future beings if humanity settles the Milky Way at carrying capacity or 10^{45} if digital life is included. Their arguments regarding strong longtermism also consider more conservative estimates: a *restricted estimate* of 10^{14} if civilization exists on Earth for 1 million years, using the average lifespan of mammalian species as a reference class, at a carrying capacity of 1 trillion lives per century, and a *low estimate* of 10^{18} if Earthbound civilization includes digital life or has a very small chance of expanding to the solar system. See also Roser (2022) (providing further estimates).

[28] As discussed in Martínez and Winter (2024), in terms of justifying legal longtermism, the extent to which the different premises must be true arguably depends on the extent to which the other premises turn out to be true. For example, the greater in size the future turns out to be, the more one can be confident in longtermism despite less feasible, predictable ways of influencing the long-term future. Conversely, the more feasible and predictable one believes it is to influence the long-term future, the smaller the number of future individuals required for one to conclude that longtermism is true. Note that the interrelation of these premises/assumptions influences not only confidence in longtermism, but also confidence in weaker versus stronger forms of longtermism; that is, depending on such calculations, one might conclude that law should be not only 'particularly' but rather 'primarily' concerned with ensuring that the long-run future goes well.

artificial intelligence. This agreement rate was lower than for any other type of risk surveyed on.[29]

Similarly, although longtermism often emphasizes the importance of law in governing emerging technologies (Winter et al. 2021), the field of law and technology is not considered normatively or descriptively central by U.S. law professors. Law and technology was the third lowest ranked area of law in terms of descriptive centrality in the U.S. Law Professor survey, and although its mean normative centrality rating was higher, it was still lower than 5 out of 10, indicating that U.S. law professors, on average, do not believe it should be central to the legal academy. Similarly, international law—recently speculated by longtermists to be a neglected area in the U.S. legal academy (Winter et al. 2021)—had a mean normative centrality rating comparable to that of local government law, an area not typically considered important by legal longtermism. These results might indicate a rejection of longtermism in favor of more near-term and traditional areas of law. However, given that the majority of participants displayed a preference for protecting the future (as indicated by endorsing personhood for those living in the near future), these results might also reflect a disparate evaluation of how to protect the long-term future. In either case, they serve as further evidence of a mismatch between the priorities of those within and those outside the field of legal longtermism.

4.2 Implications for the validity of legal longtermism

There is a burgeoning literature in the area of experimental jurisprudence dedicated to advancing philosophical, doctrinal, and policy arguments on the basis of experimental results (Prochownik 2021; Tobia 2022), including in the context of providing legal protection to the long-term future (Martínez and Winter 2024). To that end, in addition to the descriptive psychological and sociological contributions of uncovering people's views on legal longtermism, some may also view the results as having normative weight in determining (i) the appropriate level and form of legal protection for future generations *independent* of existing legal doctrine; and (ii) the appropriate level and form of legal protection for future generations according to *existing* legal doctrine.

With regard to (i), let us consider the basic abstract principle of legal longtermism that future people ought to count under the law. Within the aforementioned experimental jurisprudence literature, there is considerable debate regarding to what degree and how lay judgments—as opposed to legal expert judgments—should inform or dictate such questions of legal philosophy and policy, depending largely on the degree to which one views law through a democratic (as opposed to, say, technocratic) lens (Martínez and Winter 2024). The fact that experts and laypeople rated the desired level of legal protection for humans living in the far future as two-and-a-half to three times as high as the perceived current level, as well as the fact that the difference between the desired and perceived current level of protection was higher than any other group, arguably implies (through both a

[29] As another example, the legal community identified climate change as the most promising risk for law to address, while longtermist and legal longtermist scholarship has typically identified climate change as one of the cause areas, though not the leading one. For a thorough discussion on longtermism and climate change, see Halstead (2022).

democratic and a technocratic lens) that the existing legal institutions should be reformed so as to increase protection of humans living in the far future well beyond the current level afforded to them.

In terms of the principle that the law can protect the long-term future, consider that participants in the Global Law Professor Survey by and large agreed that there were predictable, feasible mechanisms through which the law could influence the long-term future. Insofar as law professors are experts on the potential long-term effects of law,[30] it follows that their endorsement would strengthen the same empirical premise underlying legal longtermism (i.e. that law can protect the long-term future), which in turn would provide some evidentiary and normative weight to legal longtermism.

With regard to (ii), determining the appropriate level and form of legal protection for future generations according to *existing* legal doctrine, recall that the majority of participants in the Global Law Professor Survey also endorsed standing for humans living in the near and far future in at least some possible cases. Insofar as legal academic opinion reflects or is indicative of legal doctrine as it is or ought to be interpreted, the fact that the majority of legal academics believed there to be a reasonable *legal* basis for granting standing to future generations suggests that according to at least one area of legal doctrine, future generations ought to be provided more legal protection than they are currently being granted based on existing legal doctrine. Similar reasoning applies to the U.S. Law Professor Survey, in which the majority of participants endorsed personhood for humans who will be born within 50 years in at least some cases. That is, insofar as U.S. law professors can be considered experts in legal personhood in the United States legal system, then the fact that the majority of U.S. law professors endorse personhood for some subset of future humans should provide some normative weight in favor of granting future humans personhood in at least some cases under U.S. legal doctrine.[31]

4.3 Implications for applied legal longtermism

If advocates of philosophical and legal longtermism believe that the basic abstract principles of legal longtermism are true, we would expect longtermists to want to convince others of those principles. After all, many legal scholars have postulated that both the creation and the application of the law are sensitive to—and perhaps even determined by—the will of the people. According to this view, a legal provision theoretically granting party X certain privilege Y will only be (i) passed by a legislature, (ii) interpreted as such by a judge in a relevant judicial decision, and/or (iii) commensurately enforced as such insofar as a sufficient proportion of the public is in favor of it being interpreted as such (see e.g. Post and

[30] Some might doubt the validity of this claim by arguing that although legal academics are experts in law, they are not experts in forecasting the impact of law. As mentioned in the original manuscript of the Global Law Professor Survey, future work (potentially surveying forecasting experts, or groups consisting of both forecasters and legal scholars) could help resolve this uncertainty.

[31] Note that the original prompt displayed by participants in the U.S. Law Professor Survey and U.S. Layperson Survey was: 'Insofar as the law should protect the interests of "persons", which of the following groups contains at least some "persons?"' Under one interpretation of people's responses, participants rated whether they believed the law should extend personhood to future generations beyond existing legal doctrine as opposed to merely recognizing them as persons according to existing legal doctrine. If so, one might argue that these responses provide evidence in favor of legal longtermism at the philosophical level as opposed to the doctrinal level.

Siegel 2007; Bliss 2022; but see Rodriguez Ferrere 2022). Thus, to the extent that the public does not already support granting certain privileges to future generations, this would dictate in favor of first convincing them of those principles.

However, the data presented in this chapter suggests that such convincing may not be as necessary as previously assumed, given that many, if not most, already agree with these principles. The fact that (legal) longtermism is as yet a niche approach may not be due to its counterintuitiveness but could well be explained by the fact that most people simply have not heard about it.[32] In other words, the abstract principles underlying longtermism may be intuitively appealing when asked about, despite not coming to mind easily. Consequently, one possible takeaway is that the goal of applied legal longtermism should not be so much to convince people of the validity of the abstract principles of legal longtermism but rather to raise awareness of those principles and transform those principles into concrete action and, ultimately, effective legal policy.

Future empirical work could seek to determine how best to enshrine these principles in law to protect the long-term future. While traditional approaches to evaluate the efficiency of laws and policies, such as cost-benefit analysis, are based on the assumption that humans are purely rational actors, previous work has identified various cognitive biases and heuristics that could interfere with our ability to reason about the long-term future, particularly with regard to existential and other catastrophic risks (Yudkowsky 2008; Schubert, Caviola, and Faber 2019). Since law is ultimately made, interpreted, and applied by humans, future laws, policies, and institutions should be designed to account for these biases. For example, given that (i) humans, including judges, have been found to have difficulty in reasoning about low-probability scenarios (Gatowski et al. 2001), and (ii) many of the most severe risks facing future generations involve low-probability scenarios, it follows that (iii) legislation aimed at mitigating existential risk ought to appropriately account for these limitations, such as by specifying the precise probability range to which the provision is to be applied as opposed to using more open-ended language (see Martínez and Winter 2022).

5 Conclusion

This chapter has reviewed four recent empirical studies indicating that, despite the recency of legal longtermism, the basic abstract principles underlying the theory have more resonance with legal experts and laypeople than existing laws and policies suggest. Although people do not display an equivalent preference for protecting the far future as for protecting the near-term future or present, the evidence does suggest that legal experts and laypeople across several major demographic subgroups—including political affiliation, culture, and gender—believe the law can and should protect the long-term future more than it does currently. This chapter has also discussed the implications of these results from both a theoretical and an applied standpoint. From a legal theory perspective, the fact that legal experts and laypeople largely agree that law should take seriously the interests or well-being of future people supports the validity of legal longtermism at both the philosophical and the doctrinal level. From an applied standpoint, the results suggest that the goal of applied legal

[32] This interpretation is consistent with previous literature regarding EA values more generally, finding that '[m]ost students who would agree with EA ideas haven't heard of EA yet' (Caviola, Morrissey, and Lewis 2022).

longtermism should perhaps not be so much to convince people of the validity of such abstract principles of legal longtermism beyond raising awareness, but to determine how best to enshrine those principles into concrete and effective law and policy.

References

Araújo, R., and Koessler, L. (2021), 'The Rise of the Constitutional Protection of Future Generations', Working Paper No. 7-2021 (Legal Priorities Project).

Baier, A. (1981), 'The Rights of Past and Future Persons', in E. Partridge (ed.), *Responsibilities to Future Generations: Environmental Ethics* (Prometheus Books), 171–83.

Baude, W. (2015), 'Is Originalism Our Law', in *Columbia Law Review*, 115/8: 2349–2408.

Baude, W. and Sachs, S. E. (2019), 'Grounding Originalism', in *Northwestern University Law Review*, 113/6: 1455–91.

Bliss, J. (2022), 'Existential Advocacy', Working Paper No. 4-2022 (Legal Priorities Project).

Bogojević, S. (2020), 'Human Rights of Minors and Future Generations: Global Trends and EU Environmental Law Particularities', in *Review of European, Comparative & International Environmental Law* 29/2: 191–200.

Bourget, D. and Chalmers, D. J. (2021), 'Philosophers on Philosophy: The 2020 PhilPapers Survey', https://philpapers.org/rec/BOUPOP-3 (accessed 14 January 2025).

Brand, S. (2000), 'Taking the Long View', in *Time*, https://content.time.com/time/subscriber/article/0,33009,996757,00.html (accessed 14 January 2025).

Caviola, L., Morrissey, E. S., and Lewis, J. (2022), 'Most Students Who Would Agree with EA Ideas Haven't Heard of EA Yet (Results of a Large-Scale Survey)', *EA Forum*, https://forum.effectivealtruism.org/posts/mNRNWkFBZ2K6SHD8a/most-students-who-would-agree-with-ea-ideas-haven-t-heard-of (accessed 14 January 2025).

Gaba, J. M. (1999), 'Environmental Ethics and Our Moral Relationship to Future Generations: Future Rights and Present Virtue' in *Columbia Journal of Environmental Law* 24: 249–88.

Gatowski, S. I., Dobbin, S. A., Richardson, J. T., Ginsburg, G. P., Merlino, M. L., and Dahir, V. (2001), 'Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-*Daubert* World', in *Law and Human Behavior* 25/5: 433–58.

Greaves, H. (2016), 'Cluelessness', in *Proceedings of the Aristotelian Society* 116/3: 311–39.

Greaves, H., and MacAskill, W. (2021), 'The Case for Strong Longtermism', GPI Working Paper No. 5-2021 (Global Priorities Institute, Oxford University).

González-Ricoy, I. and Gosseries, A. (eds.) (2016), *Institutions for Future Generations* (Oxford University Press).

Halstead, J. (2022), 'Climate Change & Longtermism', Supplementary Material in MacAskill, W., *What We Owe the Future* (Basic Books), https://drive.google.com/file/d/14od25qdb4sdDoXVDMoiSrTwuzYAMSpxK/view

Hart, H. (1961), *The Concept of Law* (Clarendon Press).

Hilton, B. (2022), 'Preventing an AI-Related Catastrophe', *80,000 Hours*, https://80000hours.org/problem-profiles/artificial-intelligence/ (accessed 14 January 2025).

John, T. M., and MacAskill, W. (2022), 'Longtermist Institutional Reform', in N. Cargill and John, T. M. (eds.), *The Long View: Essays on Policy, Philanthropy, and the Long-term Future* (FIRST).

Jones, N., O'Brien, M., and Ryan, T. (2018), 'Representation of Future Generations in United Kingdom Policy-Making', in *Futures* 102: 153–63.

Kannan, V. (2021), 'The Fundamental Problem with Longtermism', *Medium*, https://medium.com/@venky.physics/the-fundamental-problem-with-longtermism-33c9cfbbe7a5 (accessed 14 January 2025).

Karnofsky, H. (2016), 'Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity', *Open Philanthropy*, https://www.openphilanthropy.org/research/potential-risks-from-advanced-artificial-intelligence-the-philanthropic-opportunity/ (accessed 14 January 2025).

Lewis-Kraus, G. (2022), 'The Reluctant Prophet of Effective Altruism', in *The New Yorker*, https://www.newyorker.com/magazine/2022/08/15/the-reluctant-prophet-of-effective-altruism (accessed 14 January 2025).

MacAskill, W. (2015), *Doing Good Better: Effective Altruism and How You Can Make a Difference* (Random House).

MacAskill, W. (2017), 'Effective Eltruism: Introduction' in *Essays in Philosophy*, 18/1: 1–5.

MacAskill, W. (2022), *What We Owe The Future* (Basic Books).
Mansfield, M. (1955), 'Nuclear Weapons and the Future of Man', *Mike Mansfield Speeches, Statements and Interviews*. https://scholarworks.umt.edu/mansfield_speeches/165/
Martínez, E. and Tobia, K. (2023), 'What Do Law Professors Believe about Law and the Legal Academy? An Empirical Inquiry', Georgetown Law Journal 112, 111–189.
Martínez, E. and Winter, C. (2021a), 'Protecting Future Generations: A Global Survey of Legal Academics', Working Paper No. 1-2021 (Legal Priorities Project).
Martínez, E. and Winter, C. (2021b), 'Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection', in *Frontiers in Robotics and AI* 8: 788355.
Martínez, E. and Winter, C. (2022), 'Ordinary Meaning of Existential Risk', Working Paper No. 7-2022 (Legal Priorities Project).
Martínez, E. and Winter, C. (2024), 'Experimental Longtermist Jurisprudence', in S. Magen and K. Prochownik (eds.), *Advances in Experimental Philosophy of Law* (Bloomsbury Academic).
Martínez, E. and Winter, C. (forthcoming), 'Cross-Cultural Perceptions of Rights for Future Generations', in K. Tobia (ed.), *Cambridge Handbook of Experimental Jurisprudence* (Cambridge University Press).
Martínez, E. and Winter, C. (n.d.), 'Automating the Judiciary: A Global Survey of Legal Academics' (unpublished manuscript, on file with authors).
Minors Oposa. (1993), *Juan Antonio Oposa et al. v. the Honourable Fulgencio Factoran Jr, Secretary of the Department of the Environment and Natural Resources et al., Supreme Court of the Philippines, 30 June 1993, G.R. No 101083 (224 S.C.R.A. 792).*
Moss, D. (2020), 'EA Survey 2019 Series: How Many People Are There in the EA Community?', *EA Forum*, https://forum.effectivealtruism.org/posts/zQRHAFKGWcXXicYMo/ea-survey-2019-series-how-many-people-are-there-in-the-ea (accessed 14 January 2025).
Moss, D. (2021), 'EA Survey 2020: Engagement', *EA Forum*, https://forum.effectivealtruism.org/s/YLudF7wvkjALvAgni/p/4xczoALF6adpQk3TN (accessed 14 January 2025).
Moynihan, T. (2020), *X-Risk: How Humanity Discovered Its Own Extinction* (Urbanomic).
Neumann, J. von (1955), 'Can We Survive Technology?', in *Fortune* 51/6, 106–108, 151–52.
Olawuyi, D. (2021), 'Embedding Intergenerational Justice across Government: Regional Trends in Africa', in M-C. Cordonier Segger, M. Szabó, and A. R. Harrington (eds.), *Intergenerational Justice in Sustainable Development Treaty Implementation: Advancing Future Generations Rights through National Institutions* (Cambridge University Press), 638–55.
Open Philanthropy. (2023), 'Potential Risks from Advanced Artificial Intelligence', https://www.openphilanthropy.org/focus/potential-risks-advanced-ai/ (accessed 14 January 2025).
Ord, T. (2018), 'Fireside Chat with Toby Ord', *EA Global*, https://www.effectivealtruism.org/articles/ea-global-2018-toby-ord-fireside-chat (accessed 14 January 2025).
Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Hachette).
Parfit, D. (1984), *Reasons and Persons* (Oxford University Press).
Pellegrino, G. (2017), 'Effective Altruism and the Altruistic Repugnant Conclusion', in *Essays in Philosophy* 18/1: 44–67.
Post, R. and Siegel, R. (2007), '*Roe* Rage: Democratic Constitutionalism and Backlash', in *Harvard Civil Rights-Civil Liberties Law Review* 42: 373–433.
Prochownik, K. M. (2021), 'The Experimental Philosophy of Law: New Ways, Old Questions, and How Not to Get Lost', in *Philosophy Compass* 16/12: e12791.
Rabab Ali v Pakistan. (2016), Constitutional Petition / 1 of 2016, Supreme Court of Pakistan, April 2016.
Rodriguez Ferrere, M. B. (2022), 'Animal Welfare Underenforcement as a Rule of Law Problem', in *Animals (Basel)* 12/11: 1411.
Roser, M. (2022), 'Longtermism: The Future Is Vast—What Does This Mean for Our Own Life?', *Our World in Data*, https://ourworldindata.org/longtermism (accessed 14 January 2025).
Russell, B. and Einstein, A. (1955), 'The Russell-Einstein Manifesto', *Atomic Heritage Foundation*, https://ahf.nuclearmuseum.org/ahf/key-documents/russell-einstein-manifesto/
Sagan, C. (1994), *Pale Blue Dot* (Random House).
Sahoutara, N. (2016), 'Seven-Year-Old Girl Takes on Federal, Sindh Governments', in *The Express Tribune*, https://tribune.com.pk/story/1133023/seven-year-old-girl-takes-federal-sindh-governments (accessed 14 January 2025).
Schell, J. (2000), *The Fate of the Earth and the Abolition* (Stanford University Press).
Schoch-Spana, M., Cicero, A., Adalja, A., Gronvall, G., et al. (2017), 'Global Catastrophic Biological Risks: Toward a Working Definition', in *Health Security* 15/4: 323–28.
Schubert, S., Caviola, L., and Faber, N. S. (2019), 'The Psychology of Existential Risk: Moral Judgments About Human Extinction', in *Scientific Reports* 9:15100.

Setzer, J. and Vanhala, L. C. (2019), 'Climate Change Litigation: A Review of Research on Courts and Litigants in Climate Governance', in *WIREs Climate Change* 10/3: e580.

Sidgwick, H. (1907), *The Methods of Ethics*, 7th edition (Macmillan).

Slocum, B. G. (2019), *Ordinary Meaning: A Theory of the Most Fundamental Principle of Legal Interpretation* (University of Chicago Press).

Thorstad, D. (2021), 'The Scope of Longtermism', GPI Working Paper No. 6-2021 (Global Priorities Institute).

Tobia, K. (2021), 'Law and the Cognitive Science of Ordinary Concepts', in E. Brożek, J. Hage, and N. Vincent (eds.), *Law and Mind: A Survey of Law and the Cognitive Sciences* (Cambridge University Press), 86–93.

Tobia, K. (2022), 'Experimental Jurisprudence', in *University of Chicago Law Review* 89/3: 735–802.

UNESCO. (1997), *Declaration on the Responsibilities of the Present Generations Towards Future Generations.* https://www.unesco.org/en/legal-affairs/declaration-responsibilities-present-generations-towards-future-generations

Viña, A. G. M. La and Bueta G. R. P. (2021), 'Institutions for Future Generations in Asia', in M-C. Cordonier Segger, M. Szabó, and A. R. Harrington (eds.), *Intergenerational Justice in Sustainable Development Treaty Implementation: Advancing Future Generations Rights through National Institutions* (Cambridge University Press), 671–702.

Whittlestone, J. (2022), 'The Long-Term Future', *Effectivealtruism.org*, https://www.effectivealtruism.org/articles/cause-profile-long-run-future(accessed 14 January 2025).

Winter, C., Schuett, J., Martínez, E., Van Arsdale, S., et al. (2021), *Legal Priorities Research: A Research Agenda* (Legal Priorities Project),https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3931256 (accessed 14 January 2025).

Wu, M. (2022), 'Effective Altruism: Not Effective and Not Altruistic', in *The Phoenix*, https://swarthmorephoenix.com/2022/03/31/effective-altruism-not-effective-and-not-altruistic/ (accessed 14 January 2025).

Yassif, J. (2017), 'Reducing Global Catastrophic Biological Risks', in *Health Security* 15/4: 329–30.

Yudkowsky, E. (2008), 'Cognitive Biases Potentially Affecting Judgment of Global Risks', in N. Bostrom and M. M. Ćirković (eds.), *Global Catastrophic Risks* (Oxford University Press), 91–119.

Zhuang, J., Liang, Z., Lin, T., and De Guzman, F. (2007), 'Theory and Practice in the Choice of Social Discount Rate for Cost-Benefit Analysis: A Survey', Working Paper No. 94 (Asian Development Bank, Economics and Research Department).

31

Temporal Distance Reduces Ingroup Favoritism

Stefan Schubert, Lucius Caviola, Julian Savulescu, and Nadira S. Faber

1 Introduction

Many consider moral impartiality—treating everyone the same, regardless of who they are and where they live—to be a moral ideal. But most people don't behave in line with that ideal, instead prioritizing their ingroup—e.g., their local community or their compatriots—over outgroups (Tajfel 1970; Singer 1981; Crimston et al. 2016; 2018). This is often (but not always) associated with spatial distance: we usually prioritize people close to us over more distant people (Singer 1981).

But we are not just partial with respect to where people live and which group, community, or nation they belong to, but also with respect to *at what point in time* they live. We are *temporally partial* or *presentist*: we prioritize the present over the future, everything else equal (MacAskill 2022). People discount the future just because it is the future: they value consumption and well-being less, the further into the future it takes place (Frederick, Loewenstein, and O'Donoghue 2002; Andreoni and Sprenger 2012; Freitas-Groff 2020). Many scholars find it ethically unjustified to have such a pure time preference. They argue that assigning less value to a person and their well-being because they live at a certain point in time is just as arbitrary as letting assignments of value be influenced by someone's geographical location or the color of their skin (Ramsey 1928; Cowen and Parfit 1992; Broome 2008; Greaves 2017).

In this chapter, we study how these psychological dimensions interact. Specifically, we ask how temporal distance affects people's level of ingroup favoritism, for example their partiality for their compatriots or their local community. Are people less partial in favor of groups usually seen as their ingroup when they are asked questions relating to the distant future, as opposed to the near future? Suppose that people are asked whether to donate to a globally impartial charity or a charity prioritizing their own compatriots. Would they be more inclined to donate to the globally impartial charity if the donation would go to people living in the distant future rather than the near future? Might greater temporal distance make people less inclined to favor groups usually seen as their ingroup (either because the perceived bond to their ingroup is weakened, or because they stop making the ingroup–outgroup distinction altogether)?

There are several reasons to believe that temporal distance may reduce ingroup favoritism. The first reason may be understood in terms of relative distances. Just as we tend to feel closer to currently living ingroups than to currently living outgroups, we tend to feel closer to people who are currently alive than to people who may or will come to live

Stefan Schubert, Lucius Caviola, Julian Savulescu, and Nadira S. Faber, *Temporal Distance Reduces Ingroup Favoritism*
In: *Essays on Longtermism*. Edited by: Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press.
© Stefan Schubert, Lucius Caviola, Julian Savulescu, and Nadira S. Faber 2025. DOI: 10.1093/9780191979972.003.0031

in the future. Thus, temporal distance weakens our feelings. But it may not weaken them symmetrically. Because our feelings for currently existing foreigners are already relatively weak, adding temporal distance does not—according to this line of thinking—reduce them much further. By contrast, because our feelings for currently existing compatriots are strong, there is more room for reduction, as it were. That would mean that temporal distance would have a stronger effect on our feelings for compatriots than on our feelings for non-compatriots. In this way, it would reduce ingroup favoritism and increase our cosmopolitan tendencies.

Another hypothesis says that greater temporal distance makes people more likely to take an abstract, zoomed-out perspective on the world. It may be that when we think about the distant future (or the distant past, for that matter), we are more likely to think of humanity as a whole—on what unites us as a species—and less likely to think of the things that distinguish different groups from each other. According to this line of thinking, we view ourselves from a more distant perspective (a 'pale blue dot' in the universe; Sagan 1994) when thinking about more distant timescales. This could, in turn, promote greater cosmopolitanism and a decreased level of ingroup favoritism. This hypothesis could be derived from the so-called construal level theory of psychological distance (Trope and Lieberman 2010; 2012), but could also be defended independently (see section 3).

In Studies 1–3, we test the hypothesis that people are less inclined to prioritize their ingroup when asked about the distant future as opposed to the present and the near-term future. In Study 1, we test the basic hypothesis across a broad range of time periods, from the present to 10,000 years from now. We ask participants whether they would prioritize a charity supporting their own compatriots or a globally impartial charity, varying the time periods in which the beneficiaries of the two charities are said to live. In Study 2, we test whether any effects of time on ingroup partiality are partly driven by empirical beliefs about geopolitical changes. We reasoned that if people believe that their country will cease to exist at some point, that may drive more impartial responses to questions about the distant future. In Study 3, we test whether the effect from Studies 1 and 2 holds up for other ingroups besides the nation: people's local community, and their family and its descendants. Finally, in Study 4, we study whether temporal partiality is correlated with partiality for compatriots and other ingroups on an individual level. We reasoned that that could constitute another psychological connection between a greater focus on the distant future and lowered ingroup favoritism.

Open science

Reports of all measures, manipulations, and exclusions, as well as all data and experimental materials are available for download at https://osf.io/stw8v/?view_only=38779b24c3e04bd882d979f24bf1d3d9.

Ethics statement

For all studies, relevant ethical guidelines were followed and the research was approved through University of Oxford's Central University Research Ethics Committee.

2 Studies

2.1 Study 1

In Study 1, we aimed to test our hypothesis that people are less inclined to prioritize groups usually seen as their ingroups over outgroups when they consider issues relating to the more distant future than when they consider issues relating to the nearer future. Specifically, we asked participants whether they would prioritize giving to a charity that supports their own compatriots (US-Americans) or a charity that supports people living anywhere in the world, and varied the time periods in which these people live: today, 5 years from now, and 20, 100, 500, 2,000, and 10,000 years from now. We wanted to see whether concern for humanity as a whole would grow throughout this series of progressively more distant points in time, and hypothesized that that was the case. Our study was pre-registered at AsPredicted: https://aspredicted.org/H1G_1MB.

2.1.1 Methods

We recruited 502 US-American participants online via Mechanical Turk. They received $0.30 in payment for their participation. Four were excluded for failing a comprehension check, leaving a final sample of 498 people (231 female; M_{age} = 41.1, SD_{age} = 11.3).

We employed a within-subjects design with seven conditions that were shown to participants in a randomized order. We had one dependent variable: participants were told that they could hypothetically donate $1,000 and were asked whether they would want to prioritize donating to a charity supporting US-Americans or to a charity supporting whoever needs the money the most, regardless of country. Responses were measured on a scale from 1 (Definitely the charity supporting Americans) to 7 (Definitely the impartial charity). Depending on the condition, these charities were said to support people living today, 5 years from now, 20 years from now, 100 years from now, 500 years from now, 2,000 years from now, or 10,000 years from now.

After answering these questions, participants were asked to explain the reasons for their answers to the previous questions, and given an attention check asking who the beneficiaries of the donations were. Lastly, they responded to some demographic questions.

2.1.2 Results

Using paired sample *t*-tests, we found that people were more inclined to prioritize donating to the globally impartial charity—as opposed to the charity that prioritizes US-Americans—the further into the future the beneficiaries were said to live (see Figure 31.1). In the 10,000-year condition, participants clearly supported the impartial charity (M = 5.21, SD = 2.14). There was a significant difference ($t(497) = 3.24$, $p = .001$, $d = 0.145$) compared with the 2,000-year condition (M = 5.06, SD = 2.16), which in turn significantly differed ($t(497) = 3.17$, $p = .002$, $d = 0.142$) from the 500-year condition (M = 4.92, SD = 2.12), which in turn significantly differed ($t(497) = 8.41$, $p < .001$, $d = 0.377$) from the 100-year condition (M = 4.47, SD = 2.20), which in turn significantly differed ($t(497) = 10.3$, $p < .001$, $d = 0.462$) from the 20-year condition (M = 3.85, SD = 2.22), which in turn significantly differed ($t(497) = 7.56$, $p < .001$, $d = 0.339$) from the 5-year condition (M = 3.50, SD = 2.29), which

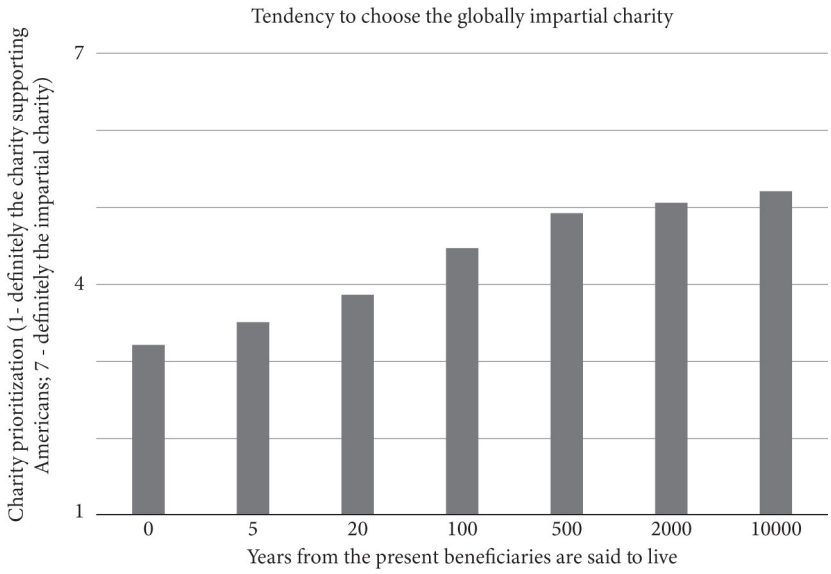


Figure 31.1 Participants' tendency to choose the globally impartial charity over the charity that only helps US-Americans increased the further into the future the beneficiaries were said to live.

in turn significantly differed ($t(497) = 5.64$, $p < .001$, $d = 0.253$) from the today condition ($M = 3.20$, $SD = 2.33$). Aggregating the seven dependent charity variables to one overall score of impartiality, we found that more politically liberal participants were significantly more likely to support the impartial charities ($r = 0.384$, $p < .001$). Similarly, less religious participants were significantly more likely to support the more impartial charities ($r = 0.152$, $p < .001$). There were no other significant demographic correlations.

2.1.3 Discussion

Study 1 gives support for our hypothesis that people are more inclined to prioritize the interests of humanity as a whole when they consider issues relating to the more distant future. In a nutshell, they are more nationalist when asked about the present and the near future, and more cosmopolitan when asked about the more distant future. Strikingly, we found differences in the full range of our studied time periods: people were more cosmopolitan about each more distant time period relative to each nearer time period. The effect was strongest when comparing different time periods in the relatively near future, whereas it grew weaker beyond 500 years. (Note also that the differences in terms of time between each time period and the next weren't the same throughout—e.g., the difference between today and 5 years from now is only 5 years, whereas the difference between 2,000 and 10,000 years from now is 8,000 years.)

2.2 Study 2

One possibility is, as discussed, that temporal distance in itself reduces ingroup favoritism, e.g., because temporal distance shrinks the relative distance between compatriots

and foreigners. According to this hypothesis, the effect of temporal distance on ingroup favoritism is driven by fundamental moral judgments. However, an alternative hypothesis is that the effect is driven by specific empirical beliefs about the future. In particular, it is possible that people believe that their country might not continue to exist in the future. Historically, many countries have ceased to exist, and it's not unreasonable to think that the same could happen with currently existing countries, such as the USA. If people don't believe that they will have any future compatriots, they may naturally be inclined to be more globally impartial with respect to the future. In Study 2, we therefore wanted to study whether people's tendency to be more globally impartial when asked about the distant future would be weakened if they were asked to assume that the USA would still exist at that point. In addition, to demonstrate robustness, we sought to replicate the effect we found in Study 1 using four additional variables besides the one about charitable giving from Study 1. These dependent variables were identification with US-Americans vs. humanity as a whole, feelings of closeness with US-Americans vs. people in distant countries, policy decisions regarding US-American vs. global interests, and responsibility to prevent catastrophes in the USA vs. catastrophes in distant countries. To simplify the study, we only included two time periods: 5 vs. 500 years. Our study was pre-registered at AsPredicted: https://aspredicted.org/PZS_W8V.

2.2.1 Methods

We recruited 605 US-American participants online via Mechanical Turk. They received $0.30 in payment for their participation. Two were excluded for failing a comprehension check, leaving a final sample of 603 people (281 female; M_{age} = 39.9, SD_{age} = 11.9).

We employed a 2x2 between-subjects design where the first factor was time (near-term, 5 years from now vs. long-term, 500 years from now) and the second factor was whether participants were asked to assume that the USA would still exist in the near-term/long-term future, or whether they were not. Participants who were asked to assume that the USA would still exist 5 [500] years from now were told: 'When you answer these questions, please assume that the USA will still exist 5 [500] years from now. Experts find it plausible that the USA will still exist 5 [500] years from now. Even if you personally find it unrealistic that the USA will still exist 5 [500] years from now, please assume that it will, for the purpose of this study.' They were then asked: 'Please indicate whether you accept the assumption that the USA will still exist 5 [500] years from now.'

We had five main dependent variables, which were presented in randomized order. One said that the USA would make a policy decision affecting the future (either 5 years from now or 500 years from now) and asked what group's interests the USA should prioritize on a scale from 1 (Definitely US interests) to 7 (Definitely the interests of humanity as a whole). Another said that participants could hypothetically donate $1,000 to charities supporting beneficiaries living 5 [500] years from now. They were asked whether they wanted to prioritize donating to a charity supporting US-Americans or to a charity supporting whoever needs the money the most regardless of country, on a scale from 1 (Definitely the charity supporting Americans) to 7 (Definitely the impartial charity). A third item asked participants whether they agreed that they are just as responsible for making sure that no great catastrophes happen to people living in distant countries as they are for making sure that no great catastrophes happen to people living in their own country, on a scale from 1 (Strongly agree) to 7 (Strongly disagree). (They were told only to consider catastrophes that

may occur 5 or 500 years from now, depending on condition.) A fourth item asked whether participants primarily identify with the USA or with humanity as a whole when they think about events 5 [500] years from now, on a scale from 1 (Definitely identify with the USA) to 7 (Definitely identify with humanity as a whole). Finally, a fifth item asked participants to indicate whether they agreed with the statement 'I feel just as close to people in distant countries as I feel towards people from the United States' (either living 5 or 500 years from now) on a scale from 1 (Strongly agree) to 7 (Strongly disagree). Participants were then asked whether they had assumed the USA would still exist 5 [500] years from now when they answered the previous questions. Subsequently, they were asked whether they had assumed that humanity would still exist 5 [500] years from now when they answered the previous questions. After answering these questions, participants were given an attention check asking which time period the questions they had been asked involved. Lastly, they were given some demographic questions.

We hypothesized, first, that across these dependent variables, participants in the long-term condition would be more favorable to humanity as a whole than participants in the near-term condition. Second, we also hypothesized that that effect would be weaker when participants were asked to assume that the USA will still exist in the long-term and near-term future, than when they were not asked to assume that.

2.2.2 Results

In line with our pre-registration, we merged all of the five dependent variables to a single score per participant (since all dependent variables significantly correlated with each other). Two of the dependent variables (about responsibility for distant and close disasters and feelings of closeness for US-Americans and people in distant countries) were reverse-scored, and then we took the average of the five variables to create an overall measure of favorability toward humanity as a whole. Performing an ANOVA (analysis of variance), we found that participants in the long-term conditions were significantly ($F(1,599) = 31.54$, $p < .001$, $\eta^2_p = .050$) more likely to be favorable to humanity as a whole than participants in the near-term conditions. We also found that participants who were asked to assume that the USA would continue to exist were significantly ($F(1,599) = 4.46$, $p = .035$, $\eta^2_p = .007$) less likely to be favorable to humanity as a whole than participants who were not, though that effect was weaker than the effect of time.

We did not find any interaction effect of time and the assumption of the USA's continued existence ($F(1,599) = 0.007$, $p = .93$, $\eta^2_p = .000$). A Tukey HSD (honestly significant difference) post-hoc test revealed that within the condition where participants were asked to assume that the USA will continue to exist, there was a significant difference ($t(599) = 3.91$, $p < .001$) between participants in the long-term condition ($M = 4.53$, $SD = 1.43$) and participants in the short-term condition ($M = 3.88$, $SD = 1.56$). The Tukey HSD post-hoc test also revealed that within the condition where participants were not asked to assume that the USA will continue to exist, there was a significant difference ($t(599) = 4.03$, $p < .001$) between participants in the long-term condition ($M = 4.80$, $SD = 1.46$) and participants in the short-term condition ($M = 4.12$, $SD = 1.39$).

Five out of 153 participants who were asked to assume that the USA would continue to exist 500 years from now rejected that assumption, as did one out of 149 participants who were asked to assume that the USA would continue to exist 5 years from now. Excluding

these participants did not make a meaningful difference to the results of our analysis. Only 20 of the 150 participants (13.3%) who had not been asked to assume that the USA would still exist 500 years from now answered that they had assumed it would not exist. More politically conservative participants were significantly more inclined toward ingroup favoritism ($r = .48$, $p < .001$), as were more religious participants ($r = .13$, $p = .001$). There were no other demographic correlations.

2.2.3 Discussion

Study 2 replicates the finding that people are more cosmopolitan about the distant future than about the near future with five dependent variables, showing that it is robust. Also, our findings lend support to the notion that this is mainly driven by fundamental moral judgments about perceived responsibilities (or similar bonds) toward compatriots and foreigners in the near and distant future. Empirical beliefs that the USA may not exist in the distant future may play a role—as they reduce the rationale for prioritizing US-Americans—but Study 2 suggests their role is smaller. First, the effect of time (distant vs. near future) was larger than that of asking participants to assume that the USA will continue to exist. Second, and crucially, we found that time (distant vs. near future) made a difference even in the condition where participants were asked to assume that the USA would continue to exist. And third, a great majority of participants who were not asked to assume that the USA will still exist 500 years from now assumed that it would. Overall, these findings suggest that even though empirical beliefs about the USA's continued existence may contribute to decreased ingroup favoritism over longer time periods, it is not the main reason for this decrease.

2.3 Study 3

Studies 1 and 2 showed that people are more likely to prioritize the interests of humanity as a whole, as opposed to those of their own nation, when they consider issues relating to the long-term future than when they consider issues relating to the near-term future. In Study 3, we wanted to see whether this tendency generalizes to other forms of partiality. We had three pre-registered hypotheses. First, we wanted to see whether people are more likely to prioritize the interests of people in their local community, as opposed to those of other compatriots, when they consider issues relating to the distant future than when they consider issues relating to the nearer future. Second, we wanted to see whether people are more likely to prioritize the interests of their own family and its descendants, as opposed to those of strangers, when they consider issues relating to the distant future than when they consider issues relating to the nearer future. We thought that it was important to see whether the results we obtained in Studies 1 and 2 specifically had to do with how people view compatriots and non-compatriots, respectively, or whether it is a more general phenomenon that concerns many different ingroup-outgroup pairs. Third, we aimed to replicate our finding that people are more likely to prioritize the interests of humanity as a whole, as opposed to those of their own nation, when they consider issues relating to the distant future than when they consider issues relating to the nearer future. Our study was pre-registered at AsPredicted: https://aspredicted.org/D6S_KBX.

2.3.1 Methods

We recruited 610 US-American participants online via Mechanical Turk. They received $0.30 in payment for their participation. Eleven were excluded for failing a comprehension check, leaving a final sample of 599 people (277 female; M_{age} = 40.7, SD_{age} = 12.7).

We employed a between-subjects design with two conditions: one long-term condition and one near-term condition. We had three main dependent variables. In each case, participants were told that they could hypothetically donate $1,000 and were asked whether they would donate it to a charity prioritizing a specific ingroup or a charity that is impartial between the ingroup and a corresponding outgroup, on a scale from 1 (Definitely the ingroup charity) to 7 (Definitely the impartial charity). It was said that the money would either help people living next year (the near-term condition) or people living 500 years from now (the long-term condition). The three pairs of potential recipients (an ingroup vs. a wider group including both the ingroup and a corresponding outgroup) of the money were 'Americans who need it' vs. 'whoever needs it the most, regardless of country'; 'people in your local village, town, or city who need it' vs. 'whatever Americans need it the most, regardless of where in the USA they live'; and 'members of your family who need it' vs. 'whoever needs it the most, including strangers'. In the long-term condition, the last of these pairs was rather 'yours and other currently living family members' descendants who need it' vs. 'whoever needs it the most, including strangers'.

After having answered these questions, participants were asked to explain the reasons for their answers to the previous questions and given an attention check asking how far into the future the recipients live. Lastly, they were given some demographic questions.

2.3.2 Results

In line with our pre-registration, we ran independent sample *t*-tests to test our hypotheses. We found, again, that people were more inclined ($t(593) = 3.91$, $p < .001$, $d = .32$) to prioritize donating to the globally impartial charity, as opposed to the charity that prioritizes US-Americans, in the long-term condition ($M = 4.32$, $SD = 2.23$) than in the near-term condition ($M = 3.59$, $SD = 2.31$). We also found that they were more inclined ($t(597) = 3.52$, $p < .001$, $d = .29$) to donate to the charity that is impartial across all US-Americans, as opposed to the charity that prioritizes their local community, in the long-term condition ($M = 4.45$, $SD = 2.20$) than in the near-term condition ($M = 3.82$, $SD = 2.16$). Lastly, we found that they were more inclined ($t(595) = 3.43$, $p < .001$, $d = .28$) to donate to the charity that is impartial between family members and strangers (including descendants), as opposed to the charity that prioritizes their family and its descendants, in the long-term condition ($M = 3.21$, $SD = 2.23$) than in the near-term condition ($M = 2.61$, $SD = 2.02$). Aggregating the three dependent variables across both conditions to one overall score of impartiality, we found that more politically liberal participants were significantly more likely to support the impartial charities ($r = 0.386$, $p < .001$), as were less religious participants ($r = 0.173$, $p < .01$), and younger participants ($r = 0.096$, $p = .002$). There were no other demographic correlations.

2.3.3 Discussion

Study 3 generalizes the results from Studies 1 and 2 and shows that they are robust across other ingroup-outgroup pairs. Previously, we found that people's tendency to prioritize their own compatriots over foreigners was weakened when faced with decisions relating to the distant future. In this study, we replicated that finding and extended it to two other

groups usually seen as ingroups: people's local community, and their family and its descendants. The fact that this tendency is stable across different ingroup-outgroup pairs suggests it is not due to some accidental feature of, e.g., how individual countries relate to other countries. Instead, it is a more general phenomenon.

2.4 Study 4

In Studies 1 to 3, we demonstrated that greater temporal distance makes people more impartial between groups usually seen as ingroups—such as their nation, their local community, or their family—and outgroups. This established a connection between temporal distance and other forms of impartiality, and raised the possibility that if people were to adopt a more long-term perspective and consider the interests of future people to a greater extent, they may also become more impartial in other regards. (But see section 3 for further discussion of this question.) In Study 4, we wanted to see whether there is also a direct connection between temporal impartiality and other forms of impartiality. Multiple streams of research going back decades have shown that different forms of partiality and bias tend to be correlated (McFarland 2010; Akrami, Ekehammar, and Bergh 2011). For instance, it has been shown that speciesism—prejudice against non-human animals—is correlated with sexism, racism, and homophobia (Caviola, Everett, and Faber 2019). These findings give us some reason to believe that temporal partiality could be correlated with ingroup favoritism. In Study 4, we therefore tested the hypothesis that temporal partiality is correlated with ingroup favoritism. We used three dependent variables—one charity variable, one feelings of closeness variable, and one moral principle variable—each of which was used to test both temporal partiality and ingroup favoritism.

2.4.1 Methods

We recruited 101 US-American participants online via Mechanical Turk. They received $0.18 in payment for their participation. Three participants were excluded for failing a comprehension check, leaving a final sample of 98 people (40 female; $M_{age} = 39.3$, $SD_{age} = 11.7$).

The study was a survey with six key dependent variables in randomized order: three relating to their degree of temporal partiality and three relating to their degree of ingroup favoritism. One of the dependent variables told participants that they could hypothetically donate $1,000 either to a charity helping people now or to a charity helping people 500 years from now (it was not specified where the beneficiaries live). They were told that if they chose the latter charity, the money would be invested and grow with time (even controlling for inflation). Participants were asked which charity they would support on a scale from 1 (Definitely the charity supporting people now) to 7 (Definitely the charity supporting people who will live 500 years from now). They were also asked an analogous question about whether they want to donate to a charity supporting US-Americans or a charity supporting whoever needs the money the most regardless of country, again on a scale from 1 (Definitely the charity supporting Americans) to 7 (Definitely the impartial charity). (It was not specified in what time-period the beneficiaries live.) Participants were also asked to what extent they agreed with the statements 'I feel just as close to people who will live 500 years from now, as I feel towards people who are alive today', 'I feel just as close to people living in distant countries, as I feel towards people from my own country', 'From a

moral perspective, people should care just as much about the well-being of people who will live 500 years from now, as they care about people who are alive today', and 'From a moral perspective, people should care just as much about the well-being of people from distant countries, as they care about people from their own country' on scales from 1 (Strongly disagree) to 7 (Strongly agree). (Again, geographical place was not specified in the questions relating to temporal partiality, and time was not specified in the questions relating to ingroup favoritism.)

After having answered these questions, participants were given an attention check where we asked in what time periods the beneficiaries of the two charities (in the first dependent variable) lived. Lastly, they were given some demographic questions.

2.4.2 Results

We merged the three dependent variables relating to degree of temporal impartiality into one measure of temporal impartiality, taking the straight average. Similarly, we merged the three dependent variables relating to degree of impartiality with respect to outgroups into one variable of outgroup impartiality, taking the straight average. We found that temporal impartiality ($M = 2.95$, $SD = 1.47$) was significantly correlated ($r = .575$, $p < .001$) with impartiality with respect to outgroups ($M = 3.96$, $SD = 1.68$). There were no significant demographic correlations (we did not test religiosity and political views this time).

2.4.3 Discussion

As expected, measures of temporal partiality were found to correlate with measures of ingroup favoritism. This matches the previously discovered pattern that different forms of partiality and bias—such as speciesism, sexism, and racism—tend to correlate. Different forms of partiality and bias may influence each other. Reduced partiality along one dimension may affect other forms of partiality. Alternatively, there may be a more general underlying common cause of different forms of partiality. Regardless, our findings suggest a connection between temporal partiality and partiality with respect to ingroups and outgroups.

3 General discussion

We found that temporal distance reduces ingroup favoritism (Studies 1–3). People tend to be less likely to favor their compatriots, members of their local community, and their family and its descendants when asked about more distant futures than when asked about the near future or the present. We also found that people who are more presentist or temporally partial were more likely to favor ingroups, and vice versa (Study 4). That further strengthens the connection between temporal partiality and partiality with respect to ingroups and outgroups.

In Study 2, we found that the observed decline in ingroup favoritism when participants are asked about the distant future is not just due to empirical beliefs about their country's (the USA's) continued existence. Participants largely retain their impartial views if asked to assume that the USA will continue to exist in the more distant future. This provides some evidence that temporal distance decreases ingroup favoritism largely because people have different moral views on ingroups and outgroups depending on whether these groups are

said to live in the near future or the more distant future. (However, we didn't study the effects of more general empirical beliefs; e.g., the belief that we are more different from more temporally distant people.)

Insofar as people's moral judgments of ingroups and outgroups change depending on which time-period the ingroups and outgroups are said to live in, what could explain that? We have not studied the relevant mechanism in detail, but it may be driven by greater temporal distance effectively shrinking geographical or social differences. People feel relatively close to currently living compatriots and relatively distant from currently living foreigners. By contrast, they may feel relatively distant to both future compatriots and future foreigners, with only a small difference between them. It may also be that temporal distance causes people to take a more zoomed-out perspective on the world, in which they focus on our common interests as a species (see section 1).

One psychological theory that could potentially explain these findings is the so-called construal level theory of psychological distance (Trope and Lieberman 2010; 2012). Construal level theory says that people mentally construe objects that are close to them along some dimension—e.g., spatially, socially, or temporally—differently from those that are distant from them. They take a more zoomed-out, abstract perspective on distant objects and focus on their central features. By contrast, people are more attentive to details and secondary features when it comes to objects that they are close to.

Research based on construal level theory has suggested that, in certain contexts, people are more cooperative and prosocial when the temporal distance is greater. For instance, it has been found that people are more likely to come to an agreement in negotiations when the temporal distance is greater (Henderson, Trope, and Carnevale 2006; De Dreu et al. 2009). People are, according to that research, more likely to zoom out from individual issues and take an integrated perspective on their overall interests from a greater temporal distance. Other research has suggested that people who are involved in a conflict are more inclined to anticipate peace and to adopt a mindset conducive to peace when they think about the more distant (as opposed to the near) future (Halevy and Berson 2022). That is potentially because peace is seen as a more abstract and aspirational state than war. Moreover, it has been found that people make harsher judgments on moral wrongs when the temporal distance is greater, since it makes them less likely to heed situational causes (Agerström and Björklund 2009). Relatedly, the same research found that temporal distance evokes stronger prosocial intuitions. All this research used construal level theory to explain their findings. Thus, it seems that there may be some broad links between a prosocial, cooperative, and moral attitude, on the one hand, and more abstract construals caused by greater temporal distance, on the other. Moreover, our hypothesis that greater temporal distance favors a more abstract perspective, where differences between ingroups and outgroups are reduced in favor of our species-wide commonalities, may fit with construal level theory. We note, however, that publication bias has been identified in replications of construal level theory effects and that taking this bias into account may cause these effects to vanish (Maier et al. 2022). It is thus important to notice that our empirical findings do not rest on construal level theory but can be assessed independently of it. Moreover, it should be noted that previous research has focused on much shorter temporal distances than we do. Nevertheless, we find the connection between our findings and previous findings based on construal level theory noteworthy.

Let us turn to limitations of our studies. Our findings are preliminary and ideally, they should be validated in a broader range of contexts. In particular, it would be useful to have

more studies of how temporal distance affects real-world behavior. More generally, some of our studies are somewhat contrived, and studies with greater degrees of realism would be useful. It can be difficult to create realistic behavioral studies of people's inclination toward the distant future, since it's hard to find interventions that reliably affect the distant future. Nevertheless, studies in that vein should be tried.

It could also be useful to study how empirical beliefs about the future could affect ingroup favoritism in greater detail. In this chapter, we only studied the effect of beliefs about the USA ceasing to exist. But as discussed, the world could also change in many other ways, from the small to the large. It may be particularly interesting to study the effects of beliefs about more radical changes. Some authors believe that the world may come to undergo a fundamental technological transformation in this or the next few centuries, fueled by artificial intelligence or other powerful general-purpose technologies (Bostrom 2009; Karnofsky 2021; MacAskill 2022). Such a technological transformation could make the world unrecognizably different, and as such render previous ingroup loyalties moot. It's therefore possible that people's degree of ingroup favoritism with respect to future beneficiaries would be further reduced if they were asked to assume that such a technological transformation would take place.

Finally, let us discuss the causal relationship between ingroup favoritism and temporal partiality. Our studies suggest that greater temporal distance reduces ingroup favoritism. That may suggest that if people became more temporally impartial, and thus more focused on future beneficiaries, then they could become more impartial between groups usually viewed as ingroups and outgroups as well. The fact that temporal partiality is correlated with partiality with respect to ingroups and outgroups is in line with that hypothesis. This could have important political consequences, since it could lead to more positive attitudes toward global cooperation. However, we cannot with any certainty say, based on our present studies, that greater temporal impartiality would have such consequences. Further research on that important question would therefore be useful.

References

Agerström, J. and Björklund, F. (2009), 'Temporal Distance and Moral Concerns: Future Morally Questionable Behavior Is Perceived As More Wrong and Evokes Stronger Prosocial Intentions', in *Basic and Applied Social Psychology* 31/1: 49–59.

Akrami, N., Ekehammar, B., and Bergh, R. (2011), 'Generalized Prejudice: Common and Specific Components', in *Psychological Science* 22/1: 57–59.

Andreoni, J. and Sprenger, C. (2012), 'Estimating Time Preferences from Convex Budgets', in *American Economic Review* 102/7: 3333–3356.

Bostrom, N. (2009), 'The Future of Humanity', in J. K. B. Olsen (ed.), *New Waves in Philosophy of Technology* (Palgrave Macmillan), 186–215.

Broome, J. (2008), 'The Ethics of Climate Change', in *Scientific American* 298: 96–102.

Caviola, L., Everett, J. A., and Faber, N. S. (2019), 'The Moral Standing of Animals: Towards a Psychology of Speciesism', in *Journal of Personality and Social Psychology* 116/6: 1011–1029.

Cowen, T. and Parfit, D. (1992), 'Against the Social Discount Rate', in P. Laslett and J. S. Fishkin (eds.) *Justice Between Age Groups and Generations* (Yale University Press), 144–161.

Crimston, C. R., Bain, P. G., Hornsey, M. J., and Bastian, B. (2016), 'Moral Expansiveness: Examining Variability in the Extension of the Moral World', in *Journal of Personality and Social Psychology* 111/4: 636–653.

Crimston, C. R., Hornsey, M. J., Bain, P. G., and Bastian, B. (2018), 'Toward a Psychology of Moral Expansiveness', in *Current Directions in Psychological Science* 27/1: 14–19.

De Dreu, C. K. W., Giacomantonio, M., Shalvi, S., and Sligte, D. (2009), 'Getting Stuck or Stepping Back: Effects of Obstacles and Construal Level in the Negotiation of Creative Solutions', in *Journal of Experimental Social Psychology* 45/3: 542–548.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002), 'Time Discounting and Time Preference: A Critical Review', in *Journal of Economic Literature* 40/2: 351–401.

Freitas-Groff, Z (2020), 'Climate Warriors and Deficit Hawks: Eliciting Intergenerational Discount Rates' (unpublished manuscript).

Greaves, H. (2017), 'Discounting for Public Policy: A Survey', in *Economics & Philosophy* 33/3: 391–439.

Halevy, N. and Berson, Y. (2022), 'Thinking about the Distant Future Promotes the Prospects of Peace: A Construal-Level Perspective on Intergroup Conflict Resolution', in *Journal of Conflict Resolution* 66/6: 1119–1143.

Henderson, M. D., Trope, Y., and Carnevale, P. J. (2006), 'Negotiation from a Near and Distant Time Perspective', in *Journal of Personality and Social Psychology* 91/4: 712–729.

Karnofsky, H. (2021), 'The "Most Important Century" Blog Post Series', *Cold Takes*, https://www.cold-takes.com/most-important-century/. Accessed 12 January 2025.

MacAskill, W. (2022), *What We Owe the Future* (Hachette).

McFarland, S. (2010), 'Authoritarianism, Social Dominance, and Other Roots of Generalized Prejudice', in *Political Psychology* 31/3:453–77.

Maier, M., Bartoš, F., Oh, M., Wagenmakers, E., Shanks, D., and Harris, A. J. L. (2022), 'Adjusting for Publication Bias Reveals That Evidence for and Size of Construal Level Theory Effects is Substantially Overestimated' in PsyArXiv. https://osf.io/preprints/psyarxiv/r8nyu

Ramsey, F. P. (1928), 'A Mathematical Theory of Saving', in *The Economic Journal* 38: 543–559.

Sagan, C. (1994), *Pale Blue Dot: A Vision of the Human Future in Space* (Random House).

Singer, P. (1981), *The Expanding Circle: Ethics and Sociobiology* (Clarendon Press).

Tajfel, H. (1970), 'Experiments in Intergroup Discrimination', in *Scientific American* 223/5: 96–103.

Trope, Y. and Liberman, N. (2010), 'Construal-Level Theory of Psychological Distance' in *Psychological Review* 117/2: 440–463.

Trope, Y. and Liberman, N. (2012), 'Construal Level Theory', in P. A. M. Van Lange, A. W. Kruglanski, and E. T. Higgins (eds.), *Handbook of Theories of Social Psychology, Volume 1* (SAGE), 118–134.

Index

For the benefit of digital users, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.

Figures are indicated by an italic *f* following the page number.

accountability 492–510
additivity 105–6
advancements 220–25
ageing 428–48
agency 384
agent-relative value 121
aggregation 40, 473
AI safety 28–29, 246–47, 275–76, 383–409
alienation 150–67
alignment 388–99, 417
altruistic willingness to pay 287–88
ambiguity 36–37
animal suffering 452
animals 449–62
anonymity 120
anti-ageing 67–69, 428–48
APS systems 385
arbitrariness 35
artificial general intelligence 410
asteroids 25
attractor state 458
authenticity 155–58
average view 443

backward induction 520
Barro-Becker model 351
beneficence 51–52, 142–43, 468
benefit ratio 19
Boltzmann brain 112–13
broadly scoped goals 423
buck-passing 533

causal diffusion 305
citizens' assembly 493
climate change 40–41, 197, 287, 378–79, 434, 451, 537
cluelessness 33–34, 154, 307
  simple 33–34
collective action 471
collective duties 471
complexification 186
conceptual frameworks 537
conformist transmission 238–39
conservatism 100–1
construal level 577
content-based transmission 241
contingency 189
contractualism 127–28
control challenge to longtermism 306
convergent evolution 190
cooperation 243–45
cost-benefit analysis 273
cryonics 69–70
cultural evolution 238–53
  cumulative 241–42
cumulative culture 192

deception 410–27
decimation-diversification hypothesis 189
deep learning 412
demandingness 41
democracy 480–81
demography 343–61, 362–82
depopulation 343–61, 362–82, 435–38
differential fitness 239
digital minds 459
disempowerment 406, 422
diversity trumps ability theorem 482–83
duties 500

economic theory 531
egalitarianism 95–97, 476
empirical economics 531
enhancements 229–30
entrenchment 27–28
epistemic challenge to longtermism 183–85, 193–94
epistemic diffusion 307
estimation 532
ethics of future generations 467
evolutionary progressivism 185–86
existential catastrophe 405, 527
existential risk 191, 211, 246, 303, 315, 343, 374, 383–409, 466
existential risk 1
existential risk factors 287
expansive longtermism 317
explicit representation 416
externalities 373, 400

fanaticism 37–39, 116–17, 471–72
  timidity 38
fertility 343–61, 362–63
final tribunal of justice 495–96
fixed cost 435–36
forecasting 171–79, 207

freedom 477
future
  duration of 21, 184
  macroevolutionary 184
  population size 21–24
futures assembly 493–94

gains 228–29
genetic engineering 456–57
global catastrophe 273
goal misgeneralization 276
goals 414–19
GPT-3, 503

healthspan 429
heuristic 415
hinge of history hypothesis 257–71, 458, 470
horizontal transmission 239
hyperopia 297

image classification 414
impartiality 105–6, 567
implicit representation 416
imprecise credences 36, 309
improvements 246
in-group favoritism 575
incommensurability 56
inequality 440
infinite justice 497
infinity 105–25
innovation 374
institutional reform 484
institutions 467–68
instrumental convergence thesis 390, 422
integrity 162
intergenerational justice 467, 475
interpersonal morality 473–74

just world bias 431
justice 473

law of logarithmic returns 328
legitimacy 479
libertarianism 478
lock-in 246–48
longtermism
  axiological 2
  axiological strong 18, 296
  counterintuitiveness of 548–49
  deontic 2, 127–31
  deontic strong 39–41, 296
  full 465–66
  individual 18, 467
  institutional 467
  justice 474–75
  legal 547
  medium 465–66
  minimal 317
  partial 335
  patient 257
  perspective 334, 513, 529
  prudential 65–85
  strict 335
  strong 2, 17, 296, 315
  urgent 257
  weak 2
loss aversion 516

machine learning 411–14
malaria 20, 355
Malthusian model 350–51
many levers argument 326–27
mass extinction 187
maturation 429
methodological hardness 534
migration 371
mind uploading 70–80
  biological 80
model-based bias 240–41
moral circle expansion 458–59
moral patiency 406
moral progress 268
moral uncertainty 309–10
myopia 197–210, 295–314, 396

natural resources 356–57
natural selection 187
neural network 411
neutrality 50–64, 88
  greediness 51
  neutral range 55
  no difference view 56
non-identity problem 40, 147, 202, 475, 480–81
non-rivalry 435–36

opacity 398
Open Philanthropy 28
optimism-pessimism dilemma 311–12
option value 303–5

pandemics 26, 274–75
Pareto principle 120
partially observable Markov decision process 299–300
path dependence 469
patient philanthropy 29–30, 321–22
perfectionism 98–99
persistent states 24–25, 249–51
person-affecting view 122, 443
personal identity 69
  branching 75
planning 386
population ethics 31–32, 86
power-seeking 383–409
prioritarianism 95–97
procreation asymmetry 87–89
productivity 373–74

proxy goals 393
PS-alignment 389
public justification 479
punctuated equilibria 189
pure longtermist goods 273

rank-dependent utility 515
reciprocity 475
reinforcement learning 414
  with human feedback 393, 417–18
repugnant conclusion 72
resolute choice 521
resource acquisition 420
retrospective accountability 494–95
returns to scale 436
reward hacking 276
reward maximization 420
reward misspecification 417
risk aversion 31, 105–6, 511

scientific progress 439
scope insensitivity 432
search 394–96
second-species arguments 410
self-sampling assumption 260
senescence 428–48
separateness of persons 142
singleton 403
situational awareness 419
space settlement 323
Spaceguard Survey 25
specialization 374
speed-ups 225–28, 321–22, 440, 453
spike, the 366*f*
spurious correlation 418
stable totalitarianism 440, 469
stakes-sensitivity argument 39, 143, 472–73
statistical value of life 441
strategic awareness 386
structural modeling 532

temporal discounting 32, 121, 152, 217–19, 296, 309–10, 432–33, 567
time of perils hypothesis 257–71, 470
total utilitarianism 54, 345, 443
totalism 108
trajectory changes 453
transformative artificial intelligence 540

unawareness 34–35
uniformitarianism 187
utilitarianism 139
  average 94–95
  critical-level 92–93
  loss averse 518
  negative 87
  total 89–92, 213

value change 458
value lock-in 458
value of information 300–2
value of statistical life 279–80
vertical transmission 239

warning shots 403
washing-out hypothesis 24
wild animals 452
winner's curse 276–77

zero-force evolutionary law 187